Speech Enhancement Based on Statistical Modeling of Teager Energy Operated Perceptual Wavelet Packet Coefficients and Adaptive Thresholding Function

by

Md. Tauhidul Islam

MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

Department of Electrical and Electronic Engineering
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

July 2014

The thesis entitled **"Speech Enhancement Based on Statistical Modeling of Teager Energy Operated Perceptual Wavelet Packet Coefficients and Adaptive Thresholding Function"** submitted by Md. Tauhidul Islam, Student No.: 0411062259, Session: April, 2011 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING on July 19, 2014.

BOARD OF EXAMINERS

1. (Dr. Celia Shahnaz) **Chairman**
   Associate Professor (Supervisor)
   Department of Electrical and Electronic Engineering
   Bangladesh University of Engineering and Technology
   Dhaka - 1000, Bangladesh.

2. (Dr. Taifur Ahmed Chowdhury) **Member**
   Professor and Head (Ex-officio)
   Department of Electrical and Electronic Engineering
   Bangladesh University of Engineering and Technology
   Dhaka - 1000, Bangladesh.

3. (Dr. Md. Saifur Rahman) **Member**
   Professor
   Department of Electrical and Electronic Engineering
   Bangladesh University of Engineering and Technology
   Dhaka - 1000, Bangladesh.

4. (Dr. Mohammad Rakibul Islam) **Member**
   Professor (External)
   Department of Electrical and Electronic Engineering
   Islamic University of Technology (IUT)
   Board Bazar, Gazipur-1704.

CANDIDATE'S DECLARATION

I, do, hereby declare that neither this thesis nor any part of it has been submitted elsewhere for the award of any degree or diploma.

Signature of the Candidate

Md. Tauhidul Islam

Dedication

To my ever caring uncle and aunt
whose inspirations are behind my every success.

Acknowledgment

I would like to mention some people who I owe a great gratitude for the moral and technical support they provided me during my thesis progress. First of all, I would like to give my greatest gratitude to my supervisor Dr. Celia Shahnaz for all her support, encouragement, guidance and useful suggestions throughout this research work. The discussions we had and her ideas not only helped me with the completion of my thesis but also gave me a different perspective and vision for my life and my future career. Thank you for your constructive comments on my work, for being friendly and tolerant to extra meetings.

I would also like to thank the rest of the members of my thesis committee: Prof. Dr. Taifur Ahmed Chowdhury, Prof. Dr. Md. Saifur Rahman, and Prof. Dr. Mohammad Rakibul Islam, for their encouragement and insightful comments. I would like to thank the head of the department of Electrical and Electronic Engineering for allowing me to use the lab facilities, which contributed greatly in completing the work in time. I wish to express note of thanks to Dr. Shaikh Anowarul Fattah, for providing inspiration and thoughtful comments.

My acknowledgments will be incomplete without the final word of gratitude to my uncle and aunt who have been the wind beneath my wings - all the way. Thanks to my parents also. Without them I would never have come so far in pursuing my dream.

Abstract

In order to handle the practical situations of real-life applications, a speech enhancement method is needed to be capable of producing optimum results with improved overall speech quality with maximized intelligibility particularly under low levels of SNRs. For solving this open problem, this thesis presents a speech enhancement approach, where an adaptive threshold is statistically determined using the Teager energy (TE) operated perceptual wavelet packet (PWP) coefficients of noisy speech. A frame of noisy speech signal is analyzed first in PWP transform domain to obtain a set of PWP coefficients. TE operation is performed on the PWP coefficients to increase the separability between clean speech and noise coefficients. The TE operated PWP coefficients with better time and frequency resolution are then used to determine an appropriate adaptive threshold based on different statistical models, namely Gaussian, Laplace, Rayleigh, Poisson and Student t distributions. The threshold thus obtained is applied upon the PWP coefficients by employing a custom thresholding function, which is designed based on the presence of noise in the noisy speech signal. A couple of custom thresholding functions designed in this thesis can be viewed as a linear combination of the modified hard or μ-law thresholding function and the semisoft thresholding function. The enhanced speech frame is synthesized by performing the inverse PWP transform on the thresholded PWP coefficients obtained using the statistically determined threshold and the designed custom thresholding function. The final enhanced speech signal is reconstructed by using the standard overlap-and-add method. Extensive Simulations using NOIZEUS database are carried out considering the presence of car and multi-talker babble noises to evaluate the performance of the proposed method in terms of standard objective metrics and subjective listening tests. It is shown that the proposed method outperforms the reported state-of the-art methods with superior efficacy at high as well as low levels of SNRs.

Contents

List of Tables

List of Figures

List of abbreviations

DWT	Discrete Wavelet Transform
DWPT	Discrete Wavelet Packet Transform
PWPT	Perceptual Wavelet Packet Transform
TEO	Teager Energy Operator
MMSE	Minimum Mean Square Error
STSA	Short Time Spectral Amplitude
SKL	Symmetric Kullback-Liebler

Chapter 1

Introduction

In this chapter, an introduction to speech enhancement, its applications, common sources of noise that degrade speech and the different types of noise those are generally considered in speech enhancement problem is provided. The main challenges and issues related to speech enhancement that motivated us to find out a new solution, objective of this thesis work along with its organization are also described.

1.1 Fundamentals of speech enhancement

Communication via speech is one of the essential functions of human beings. Humans possess varied ways to retrieve information from the outside world or to communicate with each other and the three most important sources of information are speech, images and written text. For many purposes, speech stands out as the most efficient and convenient one. Speech not only conveys linguistic contents, but also communicates other useful information like the mood of the speaker. When speaker and listener are near to each other in a quiet environment, communication is generally easy and accurate. However, at a distance or in a noisy background, the listener's ability to understand suffers. In many speech communication systems, the quality and intelligibility of speech is of greatest importance for ease and accuracy of information exchange. The speech processing systems used to communicate or store speech is usually designed for a noise free environment but in a real-world environment, the presence of background interference in the form of additive background and channel noise drastically degrades the performance of these systems, causing inaccurate information exchange and listener fatigue. Over the years, researchers have developed a number of methods to enhance speech from the degraded speech.

Yet, due to complexities of the speech signal, restoring the desired speech signal from the mixture of speech and background noise still poses a considerable challenge in speech processing and communication system research.

1.2 Applications of speech enhancement

Speech enhancement deals with processing of noisy speech signals, aiming at improving their perception by human or their correct decoding by machines. Speech enhancement algorithms attempt to improve the performance of communication systems when their input or output signals are corrupted by noise. The presence of background noise causes the quality and intelligibility of speech to degrade. Here, the quality of speech refers how a speaker conveys an utterance and includes such attributes like naturalness and speaker recognizability. Intelligibility is concerned with what the speaker had said, that is, the meaning or information content behind the words [1]. Therefore, a noisy environment reduces the speaker and listeners ability to communicate. To reduce the impact of this problem speech enhancement can be performed. It is usually difficult to reduce noise without distorting speech and thus, the performance of speech enhancement systems is limited by the tradeoff between speech distortion and noise reduction [2].

Efforts to achieve higher quality and/or intelligibility of noisy speech may effectively end up improving performance of other speech applications, such as speech coding/compression and speech recognition, hearing aids, voice communication systems and so on. The goal of speech enhancement varies according to specific applications, such as to reduce listener fatigue, to boost the overall speech quality, to increase intelligibility and to improve the performance of the voice communication device. Hence speech enhancement is necessary to avoid the degradation of speech quality and to overcome the limitations of human auditory systems.

1.3 Common Sources Of Noise for speech degradation

For communication systems, two general objectives depend on the nature of the noise and often on the signal to noise ratio (SNR) of the distorted speech. With medium to high input SNR, reducing the noise level can produce a subjectively natural speech signals at a receiver or can obtain reliable transmission. For low SNR, the objective could be to decrease the noise level, while retaining or increasing the intelligibility and reducing the fatigue caused by heavy noise for example motor and street noise. Figure 1.1 shows the factors that affect the speech signal during transmission at various stages by different noise sources. Sources that degrade speech quality are noisy environment during acquisition, background noise, multi-speaker effect, noisy transmission channel and imperfect speech reproduction. In the transmission side, the effect of background noise are added with the desired signal and the signal from other speakers are treated as noise for the desired speaker. The signal with background noise is transmitted through the channel where the transmission channel noise is also added with the desired signal.

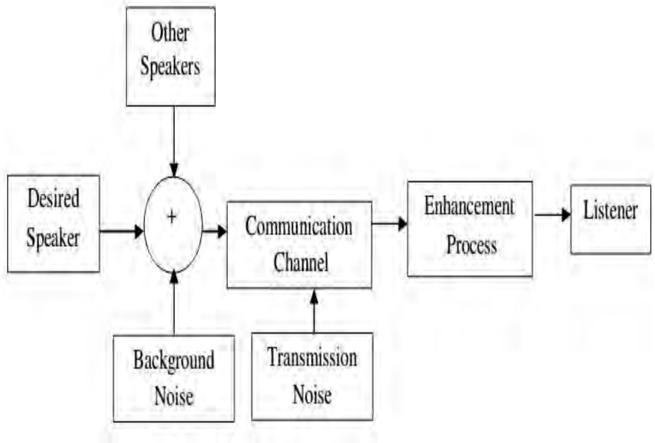


Fig. 1.1: Common sources of noise

1.4 Different types of noise

The nature of the noise is an important factor in deciding on a speech enhancement method. Therefore, a good model of noise is important for the performance of speech enhancement system and it is important to analyze how well a speech enhancement algorithm/model works with different types of noise [3]. Noise can be different based on various statistical, spectral or spatial properties. Based on the nature and properties of the noise sources, noise can be classified as additive background noise, interfering speakers (speech like noise), impulse noise, convolutive noise, and multiplicative noise. In general, it is more difficult to deal with non-stationary noise, where there is no prior knowledge available about the characteristics of noise. Since non-stationary noise is time varying, the conventional method of estimating the noise from initial intervals by assuming no speech signal is not suitable for estimation. Noise types, which are similar in temporal, frequency or spatial characteristics to speech, are also difficult to remove or attenuate. For instance, Multi talker babble retains some characteristics of speech and poses a particularly difficult problem for an algorithm intended to isolate speech signal from the background noise.

1.4.1 Additive noise

Additive noises are the noises those get added to the desired signal. In the presence of additive noise as $v[n]$, a clean speech signal $x[n]$ gets contaminated and produces noisy speech $y[n]$.

$$y[n] = x[n] + v[n]; \tag{1.1}$$

There are different types of additive noises, namely white noise and coloured noise.

White noise

White noise is defined as an uncorrelated noise process with equal power at all frequencies Fig.1.2. For example, for an audio system with a bandwidth of 10 kHz, any flat-spectrum audio noise with a bandwidth greater than 10 kHz looks like a white noise. From Fig. 1.2, it is seen that the autocorrelation function of

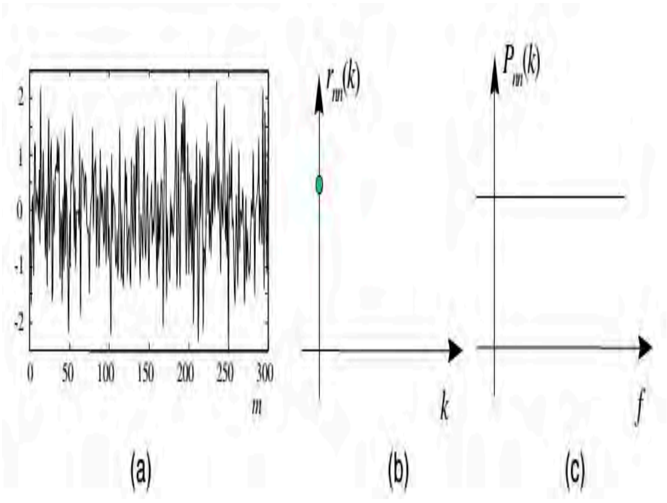


Fig. 1.2: Illustration of (a) white noise, (b) its autocorrelation, and (c) its power spectrum.

a continuous-time zero-mean white noise process with a variance of σ_2 is a delta function and white noise has a constant power spectrum.

Coloured Noise

Although the concept of white noise provides a reasonably realistic and mathematically convenient and useful approximation to some predominant noise processes encountered in telecommunication systems, many other noise processes are non-white. The term coloured noise refers to any broadband noise with a non-white spectrum. For example most audio frequency noise, such as the noise from moving cars, noise from computer fans, electric drill noise and people talking in the background, has a non-white predominantly low-frequency spectrum. Also, a white noise passing through a channel is "coloured" by the shape of the channel spectrum.

1.4.2 Multiplicative noise

In signal processing, the term multiplicative noise refers to an unwanted random signal that gets multiplied into some relevant signal during capture, transmission, or other processing. In the presence of multiplicative noise as $z[n]$, a clean speech signal $x[n]$ gets contaminated and produces noisy speech $y[n]$ as given in (1.2)

$$y[n] = x[n]z[n]; \tag{1.2}$$

1.4.3 Impulsive Noise

Impulsive noise consists of short-duration "on/off" noise pulses, caused by a variety of sources, such as switching noise, adverse channel environment in a communication system, drop-outs or surface degradation of audio recordings, clicks from computer keyboards, etc. Fig. 1.3(a) shows an ideal impulse and its frequency spectrum. In communication systems, a real impulsive-type noise has a duration that is normally more than one sample long. For example, in the context of audio signals, short-duration, sharp pulses, of up to 3 milliseconds (60 samples at a 20 kHz sampling rate) may be considered as impulsive noise. Figs. 1.3(b) and (c) illustrate two examples of short-duration pulses and their respective spectra. In a communication system, an impulsive noise originates at some point in time and space, and then propagates through the channel to the receiver. The received noise is time-dispersed and shaped by the channel, and can be considered as the channel impulse response.

In general, the characteristics of a communication channel may be linear or non-linear, stationary or time varying. Furthermore, many communication systems, in response to a large amplitude impulse, exhibit a non-linear characteristic.

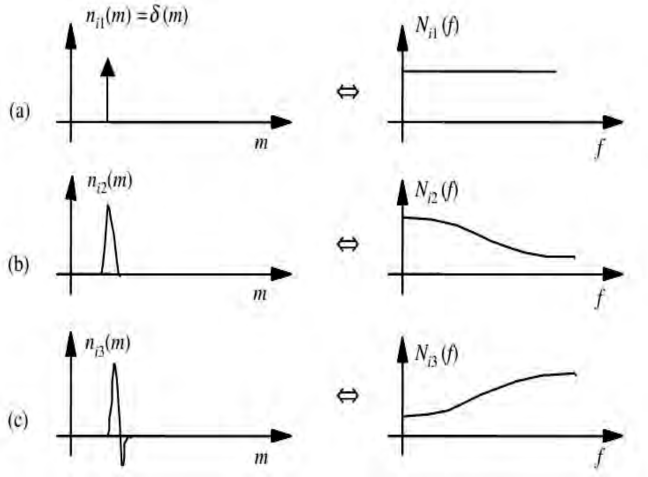


Fig. 1.3: Time and frequency characteristics of: (a) an ideal impulse, (b) and (c) shortduration pulses.

1.4.4 Stationary and non-stationary noise

In mathematics and statistics, a stationary process (or strictly stationary process or strongly stationary process) is a stochastic process whose joint probability distribution does not change when shifted in time. Consequently, parameters, such as the mean and variance, if they are present, also do not change over time and do not follow any trends. Stationarity is used as a tool in time series analysis, where the raw data is often transformed to become stationary; for example, economic data are often seasonal and/or dependent on a non-stationary price level.

Formally, let $\{X_t\}$ be a stochastic process and let $F_X(x_{t_1+\tau}, \ldots, x_{t_k+\tau})$ represent the cumulative distribution function of the joint distribution of $\{X_t\}$ at times $t_1 + \tau, \ldots, t_k + \tau$. Then, $\{X_t\}$ is said to be stationary if, for all k, for all τ, and for all $t_1, \ldots, t_k$,

$F_X(x_{t_1+\tau}, \ldots, x_{t_k+\tau}) = F_X(x_{t_1}, \ldots, x_{t_k})$. Since τ does not affect $F_X(\cdot)$,F_X is not a function of time. As an example, white noise is stationary. The sound of a cymbal clashing, if hit only once, is not stationary because the acoustic power of the clash (and hence its variance) diminishes with time. However, it would be possible to invent a stochastic process describing when the cymbal is hit, such that the overall response would form a stationary process. An example of a discrete-time stationary process where the sample space is also discrete (so that the random variable may take one of N possible values) is a Bernoulli scheme. Other examples of a discrete-time stationary process with continuous sample space include some autoregressive and moving average processes which are both subsets of the autoregressive moving average model. Models with a non-trivial autoregressive component may be either stationary or non-stationary, depending on the parameter values, and important non-stationary special cases are where unit roots exist in the model.

In this thesis, additive non-stationary coloured noises are considered.

1.5 Problem Definition

The speech signal can be acquired from single or multiple channel sensors. The multiple channel system tend to be more complex and more costly.Hence, between the two systems, the single channel systems are the most common real-time scenario algorithms, e.g., mobile communication, hearing aids etc. as usually a second channel is not available in most of such applications. Single channel speech enhancement methods can be divided mainly into three categories based on their domains of operation. Time domain methods include the subspace approach [4], frequency domain methods include the spectral subtraction [2], minimum mean square error (MMSE) estimator [5], Short-time Spectral Amplitude (STSA) estimator [6] and Wiener filtering [7], and time frequency-domain methods involve the employment of family of wavelets [8–15]. Time domain subspace method provides a tradeoff between speech distortion and residual noise but real-time processing is difficult due to heavy computation load. On the other hand, frequency domain methods provide the advantage of real-time processing with less computational load. The time-frequency domain methods, namely Universal threshold [10], WPF [9], BayesShrink [16], and SURE [17] use thresholding in the wavelet domain as a process of removing noise. The main challenge in such time-frequency domain wavelet based speech enhancement methods is adjusting the threshold value so that it can prevent distortion in enhanced speech as well as decrease residual noise. Then, by using the threshold, the designing of a thresholding function to minimize the effect of wavelet coefficients corresponding to the noise is another difficult issue. Therefore, determining an exact threshold and designing an appropriate thresholding function for noisy speech enhancement still remain as challenging tasks in the time-frequency domain.

1.6 Objective of the Thesis

The objectives of this thesis are:

i. To analyze the noisy speech signals in Perceptual Wavelet Packet (PWP) domain and perform Teager Energy (TE) Operation on PWP coefficients for better time and frequency resolution.

ii. To determine an appropriate adaptive threshold based on the statistical modeling of TE operated PWP coefficients.

iii. To develop a custom thresholding function that can operate according to the noise presence in the noisy speech signal to preserve the speech coefficients as well as to remove the noise coefficients.

iv. To investigate the performance of the proposed method in comparison with the state-of-the-art speech enhancement methods.

The outcome of this thesis is the development of a speech enhancement method based on a statistically determined accurate threshold and a custom thresholding function thus synthesizing an enhanced speech with improved quality and minimal distortion in intelligibility under high to even very low levels of SNR.

1.7 Organisation of Thesis

This thesis is organized as follows; The fundamentals and application of speech enhancement, source and types of noises are introduced in chapter 1. Chapter 2 provides a comprehensive review for the state-of-the-art speech enhancement methods. Chapter 3 describes a Gaussian model based speech enhancement method in the PWP domain. A Laplace model based speech enhancement method is discussed in chapter 4. An approach for enhancing the speech based on Rayleigh modeling is proposed in chapter 5. Chapter 6 describes a Poisson model based speech enhancement method and an approach for enhancing speech based on Student t modeling is proposed in chapter 7. Finally, concluding remarks, contribution and suggestions for future works of the thesis are highlighted in chapter 8.

Chapter 2

Literature Review

2.1 Introduction

Speech enhancement is the term used to describe algorithms or devices whose purpose is to develop some perceptual aspects of speech for the human listener or to improve the speech signal so that it may be better exploited by other speech processing algorithms. Development and widespread use of digital communication systems during the last twenty years have brought increased consideration to the role of speech enhancement in speech processing problems. Speech enhancement algorithms have been applied to problems as diverse as correction of reverberation, pitch modification, rate modification, correction of so-called ”hyperbaric” speech produced by deep-sea divers breathing a helium-oxygen mixture and correction of speech that has been distorted due to pathological problems of the speaker [1,18–21]. However, noise reduction is probably the most important and most frequently encountered speech enhancement issue. The removal of noise from degraded speech is the problem addressed in this thesis. In this chapter, a brief description of the classical methods in speech enhancement are discussed.

2.2 Time Domain Methods

One particular class of time-domain speech enhancement techniques that has gained a lot of attention is signal subspace filtering [4, 22–25].

2.2.1 Fundamentals of Subspace-based speech enhancement Method

In this approach, a nonparametric linear estimate of the unknown clean-speech signal is obtained based on a decomposition of the observed noisy signal into mutually orthogonal signal and noise subspaces. This decomposition is possible under the assumption of a low-rank linear model for speech and an uncorrelated additive (white) noise interference. Under these conditions, the energy of less correlated noise spreads over the whole observation space while the energy of the correlated speech components is concentrated in a subspace thereof. Also, the signal subspace can be recovered consistently from the noisy data. Generally speaking, noise reduction is obtained by nulling the noise subspace and by removing the noise contribution in the signal subspace. Any noise reduction technique requires assumptions about the nature of the interfering noise signal. Subspace-based speech enhancement also makes some basic assumptions about the properties of the desired signal (clean speech) as is the case in many but not all signal enhancement algorithms. Evidently, the separation of the speech and noise signals will be based on their different characteristics. Since the characteristics of the speech (and also of the noise) signal(s) are time varying, the speech enhancement procedure is performed on overlapping analysis frames.

A key assumption in all subspace-based signal enhancement algorithms is that every short-time speech vectors $[s(1), s(2), ..., s(q)]^T$ can be written as a linear combination of $p < q$ linearly independent basis function $M_i, i = 1, ..., p$, $s = M_i y$ where M_i is a (qp) matrix containing the basis functions (column-wise ordered) and y is a length-p column vector containing the weights. Both the number and the form of these basis functions will in general be time varying (frame dependent). An obvious choice forms are (damped) sinusoids motivated by the traditional sinusoidal model for speech signals. A crucial observation here is that the consecutive speech vectors s will occupy a $(p < q)$-dimensional subspace of the q-dimensional Euclidean space(p equals the signal order). Because of the time-varying nature of speech signals, the location of this signal subspace (and its dimension) will consequently be frame-dependent. The additive noise is assumed to be zero-mean, white, and uncorrelated with the speech signal. Its variance should be slowly time varying such that it can be estimated from noise only segments. Contrarily to the speech signal, consecutive noise vectors n will occupy the whole q-dimensional space.

Based on the above description of the speech and noise signals, the aforementioned q-dimensional observation space is split in two subspaces, namely a p-dimensional ($signal + noise$) subspace in which the noise interferes with the speech signal, and a *(q-p)*-dimensional subspace that contains only noise (and no speech). The speech enhancement procedure can now be summarised as follows:

1. separate the (signal+noise) subspaces from the (noise only) subspace,
2. remove the (noise-only) subspace,
3. optionally, remove the noise components in the (signal+ noise) subspace.

The first operation is straightforward for the white noise condition under consideration here, but can become complicated for the coloured noise case. The second operation is applied in all implementations of subspace-based signal enhancements, whereas the third operation is indispensable to obtain an increased noise reduction. Nevertheless, the last operation is sometimes omitted because of the introduction of speech distortion. The latter problem is inevitable since the speech and noise signals overlap in the signal subspace.

2.2.2 Algorithm of Subspace based Speech Enhancement Method

Let $s[k]$ represent the clean-speech samples and let $n[k]$ be the zero-mean, additive white noise distortion that is assumed to be uncorrelated with the clean speech. The observed noisy speech $s[k]$ is then given by

$$x[k] = s[k] + n[k] \tag{2.1}$$

Further, let R_x, R_s,and R_n be (qxq)(with q¿p)true autocorrelation matrices of x[k],s[k], and n[k], respectively. Due to the assumption of uncorrelated speech and noise, it is clear that

$$R_x = R_s + R_n \tag{2.2}$$

The eigenvalue decomposition (EVD) of R_x, R_s,and R_n can be written as follows:

$$R_s = V \Lambda V^T \tag{2.3}$$

$$R_n = V(\sigma_w^2 w)V^T \tag{2.4}$$

$$R_x = V(\Lambda + \sigma_w^2)V^T \tag{2.5}$$

with Λ diagonal matrix containing the eigenvalues λ_i, V an orthonormal matrix containing the eigenvectors v_i, σ_w^2 the noise variance, and I the identity matrix. A crucial observation here is that the eigenvectors of the noise are identical to the clean-speech eigenvectors due to the white noise assumption such that the eigenvectors of R_s can be found from the EVD of R_x in (2.7). Based on the assumption that the clean speech is confined to a $(p < q)$-dimensional subspace 2.1), we know that R_s has only p nonzero eigenvalues λ_i. If

$$\lambda_i > \sigma_w^2 \tag{2.6}$$

the noise can be separated from the speech signal, and the EVD of R_x can be rewritten as

$$R_x = V\Lambda V^T \tag{2.7}$$

if we assume that the elements λ_i of Λ are in descending order. The subscripts p and q prefer to the signal and noise subspaces, respectively. Regardless of the specific optimisation criterion, speech enhancement is now obtained by 2.1 restricting the enhanced speech to occupy solely the signal subspace by nulling its components in the noise subspace. Mathematically this enhancement procedure can be written as a filtering operation on the noisy speech vector $x = [x(1), x(2), ..., x(q)]$

$$\widehat{s} = Fx \tag{2.8}$$

with the filter matrix F given by

$$F = V_p G_p V_p^T \tag{2.9}$$

in which the (pxp) diagonal matrix G_p contains the weighting factors g_i for the first p eigenvalues of R_x, while V^T and V are known as the Karhunen Loeve transform matrix and its inverse, respectively. The filter matrix F can be rewritten as

$$F = \sum_{i=1}^{p} g_i v_i v_i^t \tag{2.10}$$

which illustrates that the filtered signal can be seen as the sum of p outputs of a "filter bank". Each filter in this filter bank is solely dependent on one eigenvector v_i and its corresponding gain factor g_i. In general, in the subspace method, a mechanism to obtain a tradeoff between speech distortion and residual noise is proposed with the cost of a heavy computational load.

2.3 Frequency Domain Methods

2.3.1 Spectral Subtraction

Spectral subtraction is the most prominent method in frequency domain [2,3,26,27]. Let $y(n) = x(n) + d(n)$ be the sampled noisy speech signal consisting of the clean speech $x(n)$ and the noise signal$d(n)$. Taking the short-time Fourier transform of $y(n)$, we get

$$Y(\omega_k) = X(\omega_k) + D(\omega_k); \tag{2.11}$$

for $\omega_k = \frac{2\pi k}{N}$ and $k = 0, 1, 2, ...N - 1$, where N is the frame length in samples. To get the short-term power spectrum of the noisy speech, we multiply $Y(\omega_k)$ in the above equation by its conjugate $Y^*(\omega_k)$. In doing so, (2.11) becomes

$$\begin{aligned} Y(\omega_k)^2 = |X(\omega_k)|^2 + |D(\omega_k)|^2 + X(\omega_k).D^*(\omega_k) \\ +X^*(\omega_k).D(\omega_k) \end{aligned} \tag{2.12}$$

Using vector to phasor conversion we get,

$$\begin{aligned} Y(\omega_k)^2 = |X(\omega_k)|^2 + |D(\omega_k)|^2 + |X(\omega_k)|e^{j\theta}.|D(\omega_k)|e^{-j\alpha} \\ +|X(\omega_k)|e^{-j\theta}.|D(\omega_k)|e^{j\alpha} \end{aligned} \tag{2.13}$$

Taking common from both sides,

$$\begin{aligned} Y(\omega_k)^2 = |X(\omega_k)|^2 + |D(\omega_k)|^2 \\ +|X(\omega_k)||D(\omega_k)|(e^{j\theta}.e^{-j\alpha} + e^{-j\theta}.e^{j\alpha}) \end{aligned} \tag{2.14}$$

We can write now,

$$Y(\omega_k)^2 = |X(\omega_k)|^2 + |D(\omega_k)|^2 + |X(\omega_k)||D(\omega_k)|(e^{jf(\theta,\alpha)}; \tag{2.15}$$

$$\begin{aligned} Y(\omega_k)^2 = |X(\omega_k)|^2 + |D(\omega_k)|^2 + |Y(\omega_k) \\ -D(\omega_k)||D(\omega_k)|(e^{jf(\theta,\alpha)} \end{aligned} \tag{2.16}$$

So at last, we get formula for the desired signal,

$$|X(\omega_k)|^2 = Y(\omega_k)^2 - |D(\omega_k)|^2$$
$$+|Y(\omega_k) - D(\omega_k)||D(\omega_k)|e^{j\angle(f(X,D)} \tag{2.17}$$

As $X(\omega_k)$ is function of $Y(\omega_k)$ and $D(\omega_k)$, we can write

$$|X(\omega_k)|^2 = Y(\omega_k)^2 - |D(\omega_k)|^2 \tag{2.18}$$
$$+|Y(\omega_k) - D(\omega_k)||D(\omega_k)|e^{j\angle(f(Y,D)} \tag{2.19}$$

We can define a gain function in the following way,

$$|\widehat{X}(\omega_k)|^2 = H^2(\omega_k)Y(\omega_k)^2 \tag{2.20}$$

where

$$H(\omega_k) = \sqrt{1 - \frac{|D(\omega_k)|^2}{|Y(\omega_k)|^2}} \tag{2.21}$$

In (2.28), the right hand side is the spectral gain function of spectral subtraction method. However, although spectral subtraction method is simple and provides a tradeoff between speech distortion and residual noise to some extent, it suffers from an artifact known as ”musical noise” having an unnatural structure that is perceptually annoying, composed of tones at random frequencies and has an increased variance. It is obvious that the effectiveness of the noise removal process is dependent on obtaining an accurate spectral estimate of the noise signal. The better the noise estimate, the lesser the residual noise content in the modified spectrum. However, since the noise spectrum cannot be directly obtained, we are forced to use an average estimate of the noise. Hence, there are some significant variations between the estimated noise spectrum and the actual noise content present in the instantaneous speech spectrum. The subtraction of these quantities results in the presence of isolated residual noise levels of large variance. These residual spectral content manifest themselves in the reconstructed time signal as varying tonal sounds resulting in a musical disturbance of an unnatural quantity. This musical noise can be even more disturbing and annoying to the listener than the distortions due to the original noise content. This and other drawbacks of the method neutralize the improvement in speech quality achieved due to the reduction in noise levels and can be more annoying than the original noise itself.

2.3.2 Minimum Mean Square Error Estimator

Minimum mean square error (MMSE) estimation of speech signals, which have been corrupted by statistically independent additive noise, is an important method in speech enhancement applications [5, 6, 28, 28, 29]. The MMSE estimator is optimal for a large class of difference distortion measures, not only the MSE measure, provided that the posterior probability density function (PDF) of the clean signal given the noisy signal is symmetric about its mean. The derivation of the MMSE estimator may be difficult, especially when complex statistical models for the signal and noise are used. In this case, the maximum a posterior (MAP) estimator of the signal, which can be efficiently calculated using the EM (expectation-maximization) algorithm, can be useful. MAP estimation is an approximate minimum average distortion estimation method for the uniform difference distortion measure. This distortion measure assigns zero distortion for estimates in the immediate neighbourhood of the clean signal, and uniform distortion for the ones outside this neighbourhood.Assuming that the MAP estimator is optimal for this non-convex distortion measure, then it is also optimal for all symmetric non-decreasing distortion measures, provided that the posterior PDF of the clean signal given the noisy signal is unimodal, symmetric about its mean, and both the distortion measure and the posterior PDF satisfy

$$\lim_{d \to \infty} d(\epsilon) P_{SIX}(\epsilon|x) = 0 \tag{2.22}$$

where $d(\epsilon)$ is the difference distortion measure, ϵ is the estimator error, and $P_{SIX}(\epsilon|x)$ is the posterior PDF of the clean signal S given the noisy signal X.

A relatively large variance of spectral coefficients is the problem of such an estimator. While adapting filter gains of the MMSE estimator, spectral outliers may emerge, that is especially difficult to avoid under noisy conditions. Unlike magnitude averaging while averaging is performed irrespective of whether the frame consists speech or noise, the MMSE estimator performs non-linear smoothing only when the SNR is low, i.e. when the frame predominantly contains noise. The residual noise present due to this technique has been observed to be colorless. The method reduces the distortions in the speech parts due to averaging.

2.3.3 Wiener Filter

Almost all of the known speech enhancement algorithms which operate in the Discrete Fourier Transform (DFT) domain assume that the real and the imaginary part of the clean speech DFT coefficients can be modeled by a Gaussian distribution. The Gaussian assumption is indeed true in the asymptotic case of large DFT frames when the span of correlation of the signal under consideration is much shorter than the DFT frame size [7]. This has been recognized, e.g., by Porter and Boll, who proposed a heuristic method to construct approximately optimal estimators from given clean speech material. In [29], different speech statistical models are investigated and based on the speech model, different MMSE estimators are obtained. Different estimators are out of the scope of this introduction, but for the exact formulas we can refer to [30]. One of the estimators based on the assumption of speech and noise being Gaussian leads to the Wiener estimator. The estimator is called a linear or Wiener filter and the formulation is [30–33],

$$\widehat{S}(k) = ES_k|X_k = \frac{\delta_s^2}{\delta_s^2 + \delta_n^2} X = \frac{\xi}{1+\xi} X \tag{2.23}$$

where δ_s^2 and δ_n^2 are the mean of $|S|^2$ and $|N|^2$.

In Wiener filter, the a priori knowledge of the speech and noise power spectra is necessary. The speech power spectrum is estimated using the estimated speech model parameters. One of the major problems of Wiener filter based methods is the requirement of obtaining clean speech statistics necessary for their implementation. Both the MMSE and Wiener estimators have a moderate computational load, but they offer no mechanism to control tradeoff between speech distortion and residual noise.

2.3.4 Short Time Spectral Amplitude Estimator

This subsection focuses on the class of speech enhancement systems that exploit the major importance of the short-time spectral amplitude (STSA) of the speech signal in its perception. In [6], a system which utilizes a minimum mean-square error (MMSE) STSA estimator is proposed. In the spectral subtraction algorithm, the STSA is estimated as the square root of the maximum likelihood (ML) estimator of each signal spectral component variance. In Wiener filtering systems, though, the STSA estimator is obtained as the modulus of the optimal minimum mean-square error (MMSE) estimator of each signal spectral component. These two STSA estimators were derived under a Gaussian assumption. As we know the spectral subtraction STSA estimator is derived from an optimal (in the ML sense) variance estimator, and the Wiener STSA estimator is derived from the optimal MMSE signal spectral estimator, so both are not optimal spectral amplitude estimators under the assumed statistical model and criterion. To derive the MMSE STSA estimator, we should know about the a priori probability distribution of the speech and noise Fourier expansion coefficients. Here we assume that the Fourier expansion Coefficients of each process can be modeled as statistically independent Gaussian random variables. Also the mean of each coefficient is assumed to be zero, since the processes involved here are assumed to have zero mean. However, due to the speech non-stationarity, the variance of each speech Fourier expansion coefficient is time-varying. The Gaussian statistical model is motivated by the central limit theorem, as each Fourier expansion coefficient is, after all, a weighted sum (or integral) of random variables resulting from the process samples. Considering the fact that a central limit theorem exists (under mild conditions) also for strongly mixing processes (i.e., in which sufficiently separated samples are weakly dependent) encourages the use of the Gaussian model in the discussed problem. The statistical independence assumption in the Gaussian model is actually correspondent to the assumption that the Fourier expansion coefficients are uncorrelated. This latter assumption is justified by the fact that the normalized correlation between different Fourier expansion coefficients approaches zero as the analysis frame length tends to infinity. In practice, a proper window (e.g., Hanning) is applied to the noisy process, which reduces the correlation between widely separated spectral components, at the expense of increasing the correlation between adjacent spectral components. This is a consequence of the wider main lobe but lower side lobes of a window function, in comparison to the rectangular window. Considering the above statistics, the MMSE estimator $\widehat{S}$ is obtained as follows

$$\widehat{S}(k) = E\{S_k|X_k\} \tag{2.24}$$

$$\widehat{S}(k) = \tau(1.5)\frac{\sqrt(\nu_k)}{\gamma_k}exp(-\frac{-\nu_k}{2})[(1+\nu_k)I_0(\frac{\nu_k}{2}) + \nu_k I_1(\frac{\nu_k}{2})]R_k \tag{2.25}$$

where $\tau(.)$ denotes the gamma function, with $\tau(1.5) = \frac{\sqrt{(\pi)}}{2}$, $I_0(.)$ and $I_1(.)$ denote the modified Bessel functions of zero and first order, respectively. ν_k is defined by:

$$\nu_k = \frac{\xi_k}{1 + \xi_k} \gamma_k \tag{2.26}$$

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)} \tag{2.27}$$

$$\gamma_k = \frac{R_k^2}{\lambda_d(k)} \tag{2.28}$$

where ξ_k and γ_k are interpreted as the a priori and a posteriori SNR, respectively. Here, such MMSE estimation of a complex exponential of the phase which does not affect the STSA estimation is done, and this constrained complex exponential estimator is found to be the complex exponential of the noisy phase. In this section the problem of estimating the a priori SNR of a spectral component in a given analysis frame is also addressed. The a priori SNR must be re-estimated in each analysis frame, due to the non-stationarity of the speech signals. Two approaches are considered here. In the first, an ML estimator of a speech spectral component variance is used. The second approach is based on a decision-directed estimation approach. Both approaches assume a prior knowledge of the noise spectral component variance. The ML estimation approach is most commonly used for estimating an unknown parameter of a given PDF, when no a priori information is available about it.

2.4 Time-Frequency domain Methods

2.4.1 Discrete Wavelet Transform based methods

Speech can be divided into two very different types of signals, namely voiced speech, such as vowels, and unvoiced speech, such as consonants. Because voiced speech is produced by the oscillation of the vocal chords it is periodic in nature. The Fourier domain is well suited for such signals, and is widely used in speech applications such as phoneme recognition. Unvoiced sounds, however, are generally not periodic in nature and the Fourier domain may not be the best way to model such signals for enhancement purposes. The success of wavelet-based signal/image enhancement has led researchers to investigate the potential of wavelet-based speech enhancement methods. Wavelet-based speech enhancement is similar to Fourier-based speech enhancement, but instead of calculating the Fourier transform of every consecutive frame, the wavelet transform is used. Thresholding speech in the wavelet domain can easily eliminate sections of speech, though, especially when enhancing the noise-like unvoiced sounds. The algorithm uses voiced/unvoiced detection to solve this problem. Unvoiced sections of speech are enhanced by only attenuating the coefficients of the highest resolution level, whereas all coefficients are attenuated with voiced sounds. Bahoara and Rouat [9] proposed a speech enhancement algorithm by using a time-adaptive threshold in a 16-subband uniform wavelet packet domain. Bahoara and Rouat reported that that their algorithm improves the global SNR more than the Ephraim-Malah MMSE STSA algorithm [6], even under heavy noise conditions.

Hu and Loizou [23] proposed a different approach which also combines short-time spectral attenuation (STSA) and wavelet-based enhancing techniques. Unlike the above-mentioned wavelet-based algorithms, which threshold the wavelet coefficients of the time signal, this algorithm enhances the log multitaper spectra [34]. The multitaper spectra have good bias and variance properties. These spectral signals are then transformed to the wavelet domain, enhanced with SureShrink and then finally inverse transformed back into the log multitaper spectral domain. Wavelet enhancing of the log multitaper spectra leads to even better (low-variance) spectral estimates. These refined spectra are then used in an STSA speech enhancement algorithm, which is a variation of Wiener filtering. The actual speech enhancement is done in the multitaper spectral domain, whereas the wavelet-based enhancement step is only used to get more refined spectral estimates, which makes this algorithm an STSA speech enhancement algorithm. Hu and Loizou showed that their algorithm has little "musical" noise and it also preserves speech quality better than the Ephraim-Malah MMSE-LSA algorithm [4].

In the conventional DWT based analysis, only scale space is decomposed, but wavelet space is not decomposed. An important shortcoming of such analysis when it is applied to the noisy speech for the purpose of enhancement is the shrinkage of the unvoiced speech frames which contain many noise-like speech components leading to a degraded speech quality.

2.4.2 Discrete Wavelet packet Transform based Methods

Unlike DWT based analysis, in Wavelet Packet (WP) based analysis, the wavelet space is also decomposed thus making the higher frequency band decomposition possible. Since, both the approximation and the detail WP coefficients are decomposed into two parts at each level of decomposition, a complete binary tree with superior frequency localisation can be achieved. Discrete Wavelet Packet Transform (DWPT) decomposes the signal into a larger number of subbands and produces a multiresolution framework that can have finer frequency resolution at high frequencies than the standard wavelet-transform [8–15].

Cohen [35] proposed an algorithm which uses a weighted Wiener filter to attenuate the coefficients of a non-uniform 84-subband redundant DWPT. The subband spacing approximates the bark frequency scale, which is a perceptual frequency scale generally used for audio compression purposes. The a priori SNR is estimated by a variation of the Ephraim Malah decision-directed estimate [4]. Compared to Fourier-based speech enhancement, the algorithm leads to better results on the segmental signal-to-noise ratio distortion measure and lower residual noise of enhanced speech.

Fu and Wan [36] proposed a method which uses Fourier-based and wavelet-based denoising techniques in a series combination. The Ephraim-Malah MMSE STSA speech enhancement algorithm [4] is used as a pre-processing step to eliminate some noise while still retaining speech quality. This enhanced speech signal is then transformed into the DWPT domain by using an 18-subband critical-band decomposition. Time and frequency-adaptive thresholds are computed for each subband and time frame by using a variation of the universal threshold. Enhancement is done with a variation of the Ephraim Malah suppression rule [4]. Fu and Wan state that combining Fourier-based and wavelet-based enhancement techniques eliminates a reasonable amount of ”musical” noise while still retaining speech quality. The algorithm also shows promising results on the segmental signal-to-noise ratio distortion measure.

2.5 Conclusion

In this chapter, a brief literature survey of the state-of-the-art speech enhancement methods are presented. All the methods have their advantages and disadvantages. In order to handle the practical situations of real life applications, a speech enhancement method, apart from providing less computational burden, is needed to be capable of producing satisfactory results with improved speech intelligibility. Although a series of successful attempt has been taken by many researchers, it is still a open challenge problem.

Chapter 3

Speech Enhancement Using Gaussian Modeling of Teager Energy Operated Perceptual Wavelet Packet Coefficients

In this chapter, speech enhancement based on Gaussian modeling of TE operated PWP coefficients is described [37]. An adaptive threshold is determined analytically using the Gaussian model of TE operated PWP coefficients and then this threshold is imposed upon the PWP coefficients of noisy speech using pdf dependent custom thresholding function which is devised as a combination of modified hard and semisoft thresholding function. Detail simulation is performed to compare the proposed method with the state-of-the art speech enhancement techniques which is added at the end of this chapter.

3.1 Proposed Method Considering Gaussian Statistical Model

The block diagram for the proposed method is shown in Fig. 3.1. It is seen from Fig. 3.1 that PWP transform is first applied to each input speech frame. Then, the PWP coefficients are subject to TE approximation with a view to determine a threshold value for performing thresholding operation in the WP domain. On using a custom thresholding function, an enhanced speech frame is obtained via inverse perceptual wavelet packet (IPWP) transform.

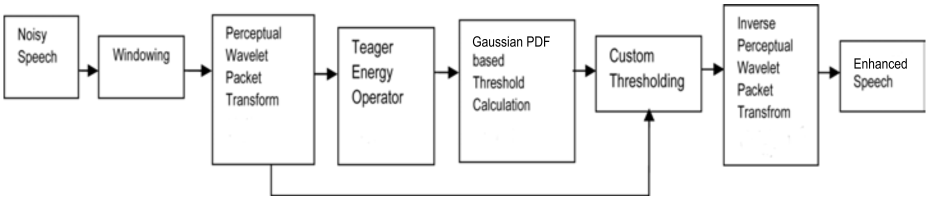

Noisy Speech
Windowing
Perceptual Wavelet Packet Transform
Teager Energy Operator
Gaussian PDF based Threshold Calculation
Custom Thresholding
Inverse Perceptual Wavelet Packet Transfrom
Enhanced Speech

Fig. 3.1: Block diagram for the proposed method

3.1.1 Perceptual Wavelet Packet Transform

The perceptual scale mel scale, named by Stevens, Volkman and Newman in 1937 is a scale of pitches judged by listeners to be equal in distance from one another. The reference point between this scale and normal frequency measurement is defined by assigning a perceptual pitch of 1000 mels to a 1000 Hz tone, 40 dB above the listener's threshold. Above about 500 Hz, larger and larger intervals are judged by listeners to produce equal pitch increments. As a result, four octaves on the hertz scale above 500 Hz are judged to comprise about two octaves on the mel scale.

Formula to convert f hertz into m mel is,

$$m = 1127log(1 + \frac{f}{700}) \tag{3.1}$$

The conversion of frequency to mel is shown in Fig.3.6.

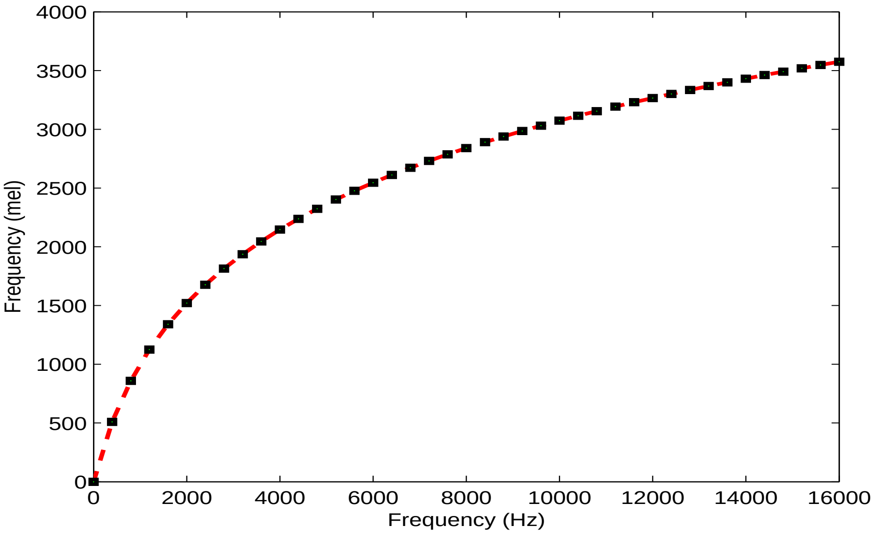


Fig. 3.2: Conversion of frequency to perceptual scale mel

The perceptual wavelet packet transform transforms the wavelet coefficients according to the frequency division of perceptual scale. The main motivation behind this transform is the ability to decompose the signal according to human auditory system. At low frequency, where human auditory system can differentiate the pitches precisely, PWPT decomposes the signal in finer bands. On the other hand, at high frequency, PWPT creates less number of bands as the human cochlea can not differentiate small differences in high frequency.

The method introduced here is based on perceptual wavelet packet decomposition. The key element of the transform is the use of the Mel warping function to determine the WPT decomposition structure based on a perceptually motivated frequency axis. we propose to decompose a wavelet packet tree into the critical bands with respect to the Mel frequency warping curve [38]. The frequency division for a perecptual wavelet packet transform is shown in the Fig. 3.3. The center frequencies of the wavelet packet transform and perceptual wavelet packet transform are shown in Table. 3.1.

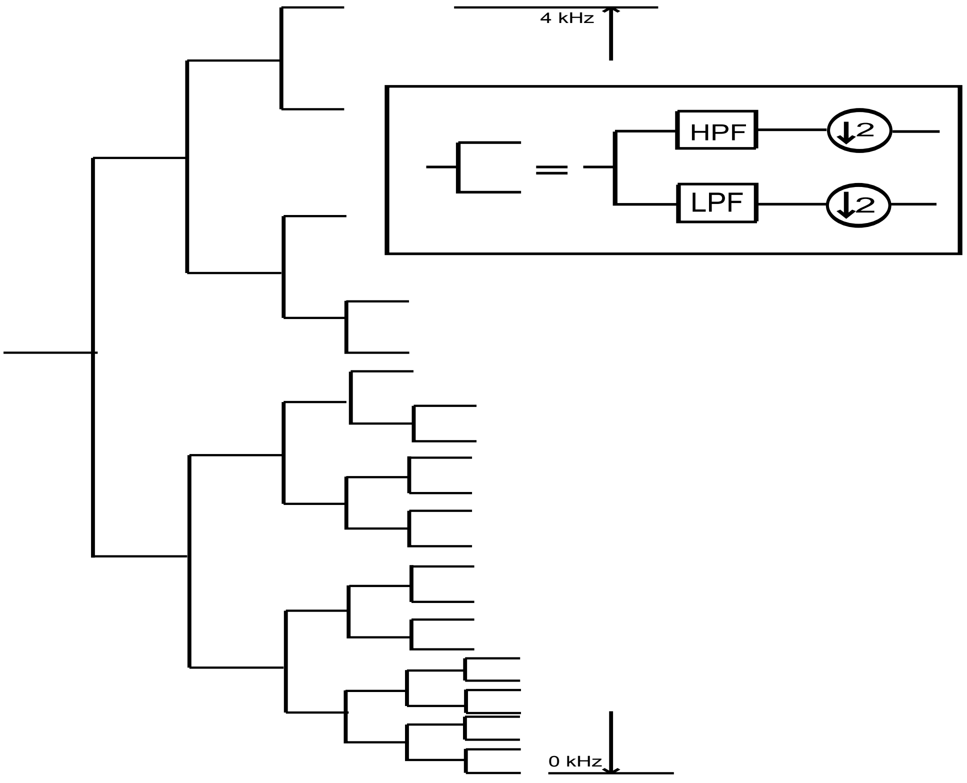


Fig. 3.3: Frequecy Structure for Perceptual Wavelet Packet Transform

The clean, noise and noisy PWP coefficients in a subband of a noisy speech frame at an SNR of 5dB is plotted in Fig. 3.4. It is seen from this figure that for most of the coefficient indices, clean and noise PWP coefficients are not separable. Based on similar analysis performed on many speech signals corrupted by different noises, it

Table 3.1: Center Frequency of WPT and PWPT

Filters	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
PWPT	28	89	154	224	300	383	472	569	674	787	910	1043	1187	1343	1512	1694	1892	2106	2338	2589	2860	3154	3472	3817
WPT	31	94	156	219	281	344	406	469	563	688	813	938	1063	1188	1313	1438	1563	1688	1875	2125	2375	2750	3250	3750

is found that the time and frequency resolution provided by PWP transform is not sufficient to separate PWP coefficients of clean speech from that of noise even at a high SNR of 5dB. Since, Teager Energy (TE) operator has better time and frequency resolution [39] it can be very useful in handling noise. Therefore, we apply discrete time TE operator on the PWP coeffcients $W_{k,m}$.

3.1.2 Teager Energy Operator

Letting $W_{k,m}$ as the m-th PWP coefficient in the k-th subband, the m-th TE operated coefficient $t_{k,m}$ corresponding to the k-th subband of the PWP transform is given by

$$t_{k,m} = T[W_{k,m}]. \tag{3.2}$$

Fig. 3.5 presents the clean, noise and noisy TE operated PWP coefficients in a subband of a noisy speech frame at the same SNR as used in fig.3.4. It is seen from this figure that at the indices where TE operated PWP coefficients of clean speech have higher values, the TE operated PWP coefficients of noise show lower values. As a result, thresholding operation on the noisy PWP coefficients needs a low threshold value thus removing the noise leaving the speech undistorted. On the contrary, at the indices, where TE operated PWP coefficients of clean speech have lower values, the TE operated PWP coefficients of noise exhibit higher values as expected. Thus thresholding the noisy speech PWP coefficients needs a higher threshold value and removes the necessary noise without speech distortion at a significant level. Therefore, TE operation on PWP coefficients is found as more capable of serving the goal of thresholding operation by reducing the noise as well as preserving the speech.

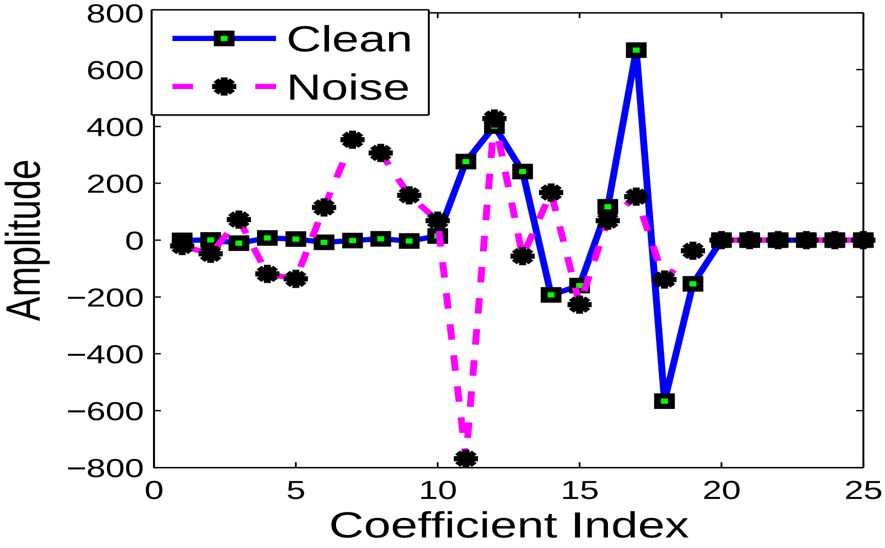


Fig. 3.4: WP Coefficients of a noisy speech subband at an SNR of 5dB

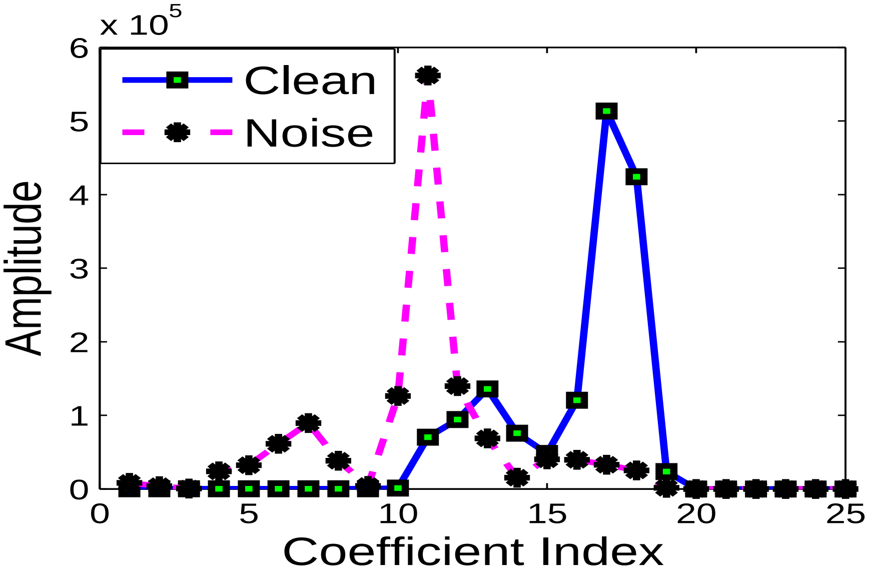


Fig. 3.5: TE Operated PWP Coefficients of a noisy speech subband at an SNR of 5dB

3.1.3 Proposed Model for TE Operated PWP Coefficients assuming Gaussian distribution

The outcome of a speech enhancement method based on the thresholding in a transform domain depends mainly on two factors, namely the threshold value and the thresholding functions. The use of a unique threshold for all the PWP subbands is not reasonable. As a crucial parameter, the threshold value in each subband is required to be adjusted very precisely so that it can prevent distortion in the enhanced speech as well as decrease annoying residual noise. By considering the probability distributions of the $t_{k,m}$ of the noisy speech, noise and clean speech, a more accurate threshold value can be obtained using a suitable pattern matching scheme or similarity measure. Since speech is a time-varying signal, it is difficult to realize the actual probability distribution function (pdf) of speech or its $t_{k,m}$. As an alternative to formulate a pdf of the of speech, we can easily formulate the histogram of its $t_{k,m}$ and can approximate the histogram by a reasonably close pdf namely Gaussian distribution. For the $t_{k,m}$s in a subband of a noisy speech frame, the empirical histogram along with the Gaussian distributions are superimposed in Fig. 3.6, 3.7 and 3.8 in presence of car noise at SNRs of -15, 0 and 15 dB. From this figure, it is obvious that Gaussian distribution fits the empirical histogram very finely. Similar analysis results are obtained for empirical histogram and Gaussian distribution of TE operated noise PWP coefficients at the same SNRs as used in Fig. 3.6, 3.7 and 3.8 and are shown in Fig. 3.9, 3.10 and 3.11.

3.1.4 Proposed Adaptive Threshold Calculation assuming Gaussian distribution

The entropy of each subband of the PWP coefficients is found different from each other. So, an entropy measure may be chosen to select a suitable threshold value adaptive to each subband. Some popular similarity measures that are related to the entropy functions are the Variational distance, the Bhattacharyya distance, the Harmonic mean, the Kullback Leibler(K-L) divergence, and the Symmetric K-L divergence. The K-L divergence is always nonnegative and zero if and only if the approximate Gaussian distribution functions of the pdf of noisy speech and that of the noise or the approximate Gaussian distribution functions of the pdf of the noisy

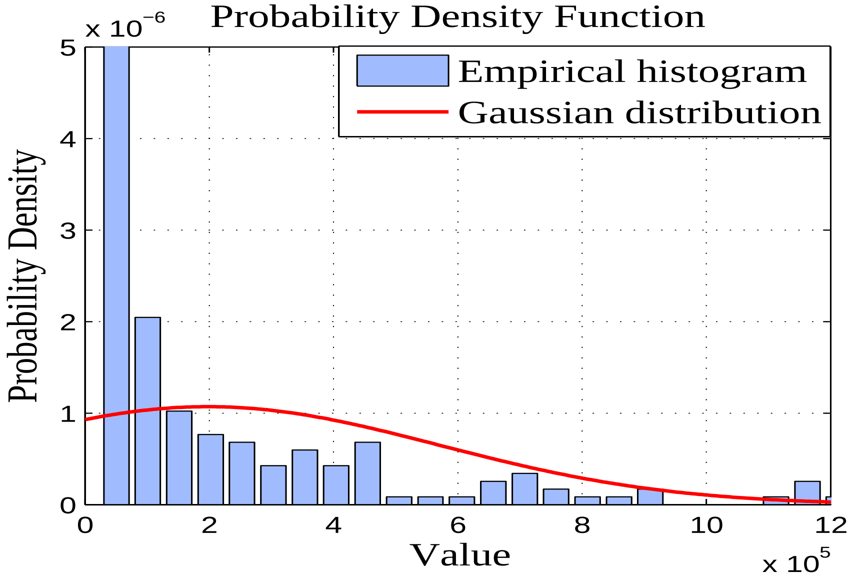


Fig. 3.6: Empirical histogram and Gaussian distribution of TE operated PWP coefficients of noisy speech at SNR of -15 dB

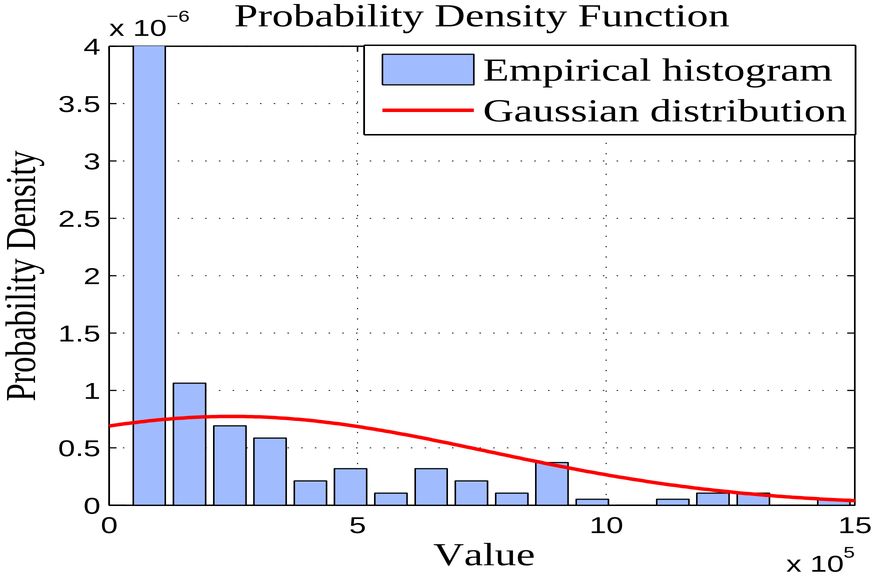


Fig. 3.7: Empirical histogram and Gaussian distribution of TE operated PWP coefficients of noisy speech at SNR of 0 dB

speech and that of the clean speech are exactly the same. In order to have a symmetric distance between the any two approximate Gaussian distribution functions as mentioned above, the Symmetric K-L divergence has been adopted in this paper. The Symmetric K-L divergence is defined as

$$SKL(p,q) = \frac{KL(p,q) + KL(q,p)}{2}, \tag{3.3}$$

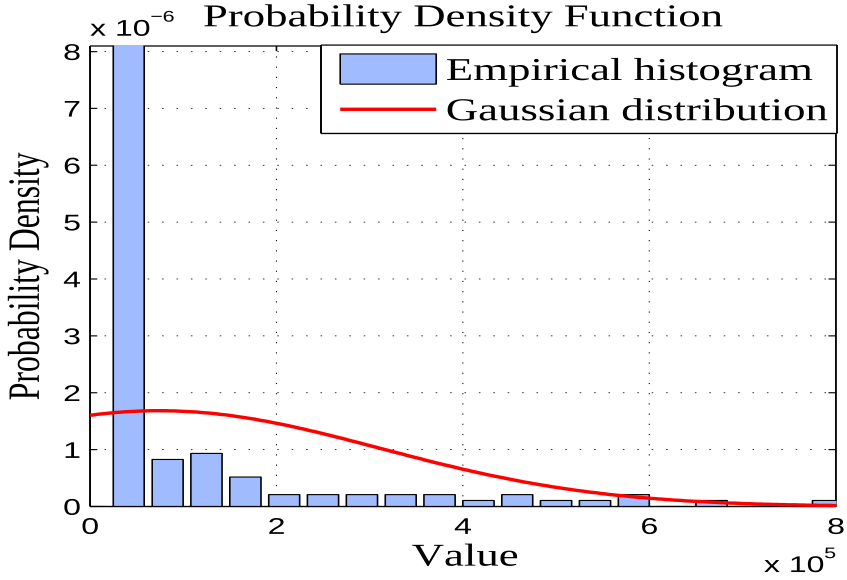


Fig. 3.8: Empirical histogram and Gaussian distribution of TE operated PWP coefficients of noisy speech at SNR of 15 dB

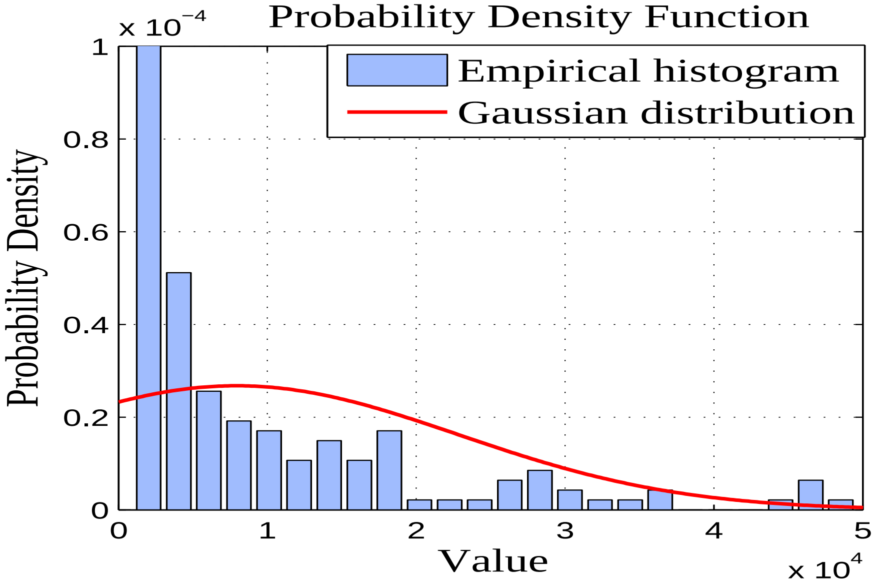


Fig. 3.9: Empirical histogram and Gaussian distribution of TE operated noise PWP coefficients at SNR of −15 dB

where p and q are the two approximate Gaussian pdfs calculated from the corresponding histograms each having N number of bins and $KL(p,q)$ is the K-L divergence given by

$$KL(p,q) = \sum_{i=1}^{N} p_i(t_{k,m}) ln \frac{p_i(t_{k,m})}{q_i(t_{k,m})}. \tag{3.4}$$

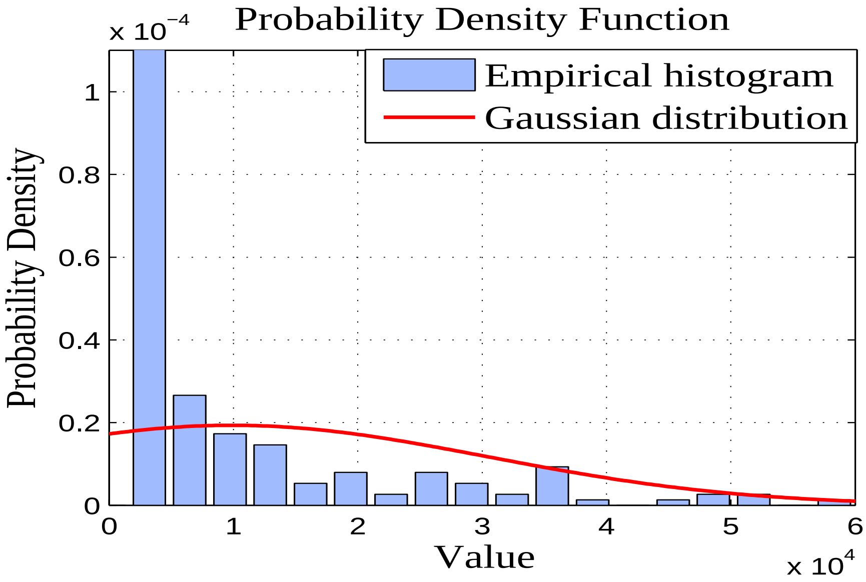


Fig. 3.10: Empirical histogram and Gaussian distribution of TE operated noise PWP coefficients at SNR of 0 dB

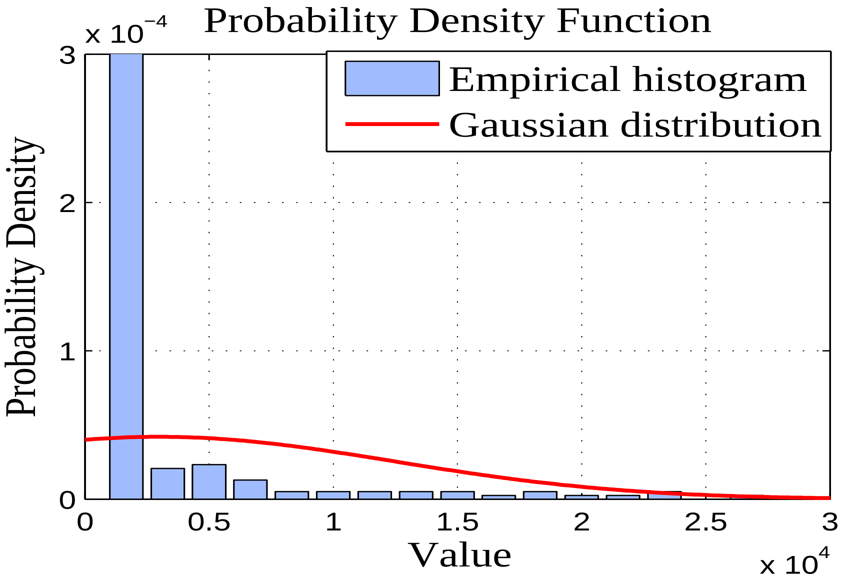


Fig. 3.11: Empirical histogram and Gaussian distribution of TE operated noise PWP coefficients at SNR of 15 dB

In (3.4), $p_i(t_{k,m})$ is the pdf of $t_{k,m}$ of noisy speech given by

$$p_i(t_{k,m}) = \frac{n_i}{N_c}, \tag{3.5}$$

where n_i is number of coefficients in i-th bin and N_c total number of coefficients in each subband. Similarly, the approximate Gaussian pdf of the $t_{k,m}$ of the noise and that of the $t_{k,m}$ of the clean speech can be estimated following (3.5) and denoted by $q_i(t_{k,m})$ and $r_i(t_{k,m})$, respectively. Below a certain value of threshold λ, the symmetric K-L divergence between $p_i(t_{k,m})$ and $q_i(t_{k,m})$ is approximately zero, i.e.,

$$SKL(p_i(t_{k,m}), q_i(t_{k,m})) \approx 0. \tag{3.6}$$

By solving the above equation, we get a value of λ following [40],

$$\lambda(k) = \frac{\sigma_n(k)}{\sqrt{\gamma_k}}\sqrt{2(\gamma_k + \gamma_k^2)} \times ln(\sqrt{1 + \frac{1}{\gamma_k}}), \tag{3.7}$$

where γ_k is segmental SNR of subband k defined as

$$\gamma_k = \frac{\sigma_r^2(k)}{\sigma_n^2(k)}. \tag{3.8}$$

In this equation, $\sigma_r^2(k)$ is the signal power at k subband and $\sigma_n^2(k)$ is the noise power at k subband.

3.1.5 Proposed Thresholding Function

We propose a pdf dependent custom thresholding function derived from the modified hard and the semisoft thresholding functions [41]. Representing $\lambda(k)$ derived from (3.7) as $\lambda_1(k)$ and letting $\lambda_2(k) = 2\lambda_1(k)$, the proposed thresholding function is developed as

$$(Y_{k,m})_{PCT} = \begin{cases} \alpha(k,m)sgn(Y_{k,m}) \times G, & \text{if } |(Y_{k,m})| < \lambda_1(k) \\ Y_{k,m}, & \text{if } |(Y_{k,m})| > \lambda_2(k), \\ (1-\alpha(k,m))\Pi_1 + \alpha(k,m)\Pi_2, & \text{otherwise,} \end{cases} \tag{3.9}$$

where

$$G = \frac{|(Y_{k,m})|^{\beta(k,m)}}{[\lambda_1(k)]^{(\beta(k,m)-1)}}, \tag{3.10}$$

$$\Pi_1 = sgn(Y_{k,m}) \times \lambda_2(k)\frac{|(Y_{k,m})| - \lambda_1(k)}{\lambda_2(k) - \lambda_1(k)}, \tag{3.11}$$

$$\Pi_2 = Y_{k,m}. \tag{3.12}$$

In (3.9), $(Y_{k,m})_{PCT}$ stands for the PWP coefficients thresholded by the proposed custom thresholding function expressed from (3.9)-(3.12) and two shape parameters of the proposed thresholding function are represented by $\alpha(k,m)$ and $\beta(k,m)$.

The comparison of the proposed custom thresholding function with the conventional modified hard and semisoft thresholding functions is shown in Fig. 3.12. In the region between λ_1 and λ_2, this figure demonstrates the flexibility of the proposed thresholding operation in a sense that it can be viewed as $(1 - \alpha(k,m))(Y_{k,m})_{SS} +$

$\alpha(k,m)(Y_{k,m})_{MH}$ which is a linear combination of the modified hard and the semisoft thresholding function. Here, $(Y_{k,m})_{MH}$ stands for the PWP coefficients thresholded by the modified hard thresholding function and $(Y_{k,m})_{SS}$ represents the PWP coefficients thresholded by the semisoft thresholding function. Unlike these functions, depending on the value of shape parameter $\alpha(k,m)$, it can be verified from (3.9) that the proposed thresholding function gets the following forms,

$$\lim_{\alpha(k,m)\to 0}(Y_{k,m})_{PCT} = (Y_{k,m})_{SS},$$
$$\lim_{\alpha(k,m)\to 1}(Y_{k,m})_{PCT} = (Y_{k,m})_{MH}.$$

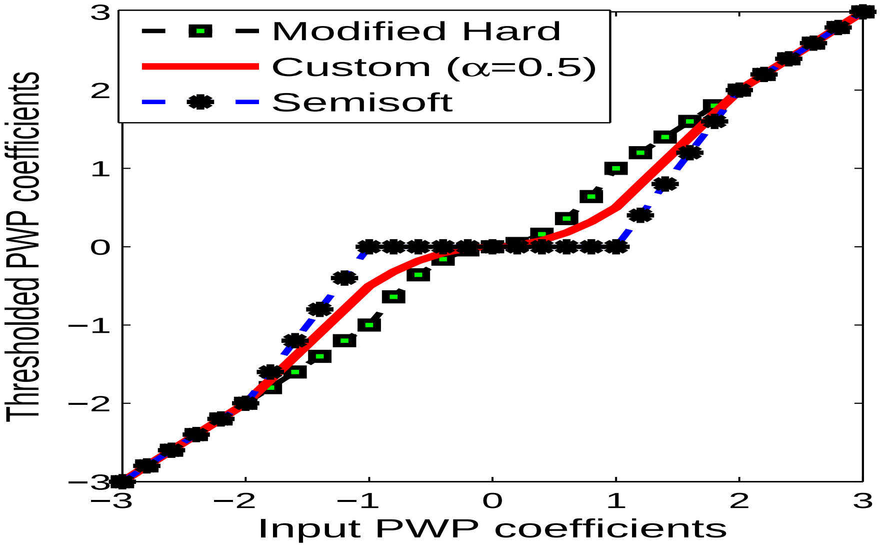


Fig. 3.12: Input Output Relation for semisoft, modified hard and proposed custom thresholding function

Effect of the Shape Parameters on the Proposed Thresholding Function

In order to realize the effect of $\alpha(k,m)$ and $\beta(k,m)$ on the proposed thresholding function, the variation of $\alpha(k,m)$ and $\beta(k,m)$ for different values of $\frac{R(k,m)}{Q(k,m)}$ are obtained using (7.11) and (7.12) and plotted in Fig.3.13. From this figure, it is seen that for a large value of $\frac{R(k,m)}{Q(k,m)}$, $\alpha(k,m)$ becomes high, i.e., close to 1 that increases the probability of $Y_{k,m}$ to be a speech coefficient. In this case, $(Y_{k,m})_{PCT}$ acts like $(Y_{k,m})_{MH}$ as expected, since if a coefficient has a high probability to be speech should not be a thresholded to zero before $\lambda_1(k)$ and should be unchanged after $\lambda_1(k)$ as done in modified hard thresholding function. It is also found from Fig.

3.13 that for a small value of $\frac{R(k,m)}{Q(k,m)}$, $\alpha(k,m)$ becomes low, i.e., close to zero and $(Y_{k,m})_{PCT} \approx (Y_{k,m})_{SS}$. It is also expected since if the probability of a PWP coefficient to be speech becomes low, it should be thresholded to zero before $\lambda_1(k)$ and thresholded to a small value upto $\lambda_2(k)$ as done in semisoft thresholding function. From Fig.3.13, it is seen that for a small value of $\frac{R(k,m)}{Q(k,m)}$, $\beta(k,m)$ gets a high value that increases the probability of $Y_{k,m}$ to be a noise coefficient. In this case, it can be seen from (3.10) that $(Y_{k,m})_{PCT}$ in (3.9) tends to zero as expected since a noise PWP coefficient should be made zero to completely remove the noise. On the other hand, for a high value of $\frac{R(k,m)}{Q(k,m)}$, $\beta(k,m)$ becomes low that decreases the probability of $Y_{k,m}$ to be a noise coefficient. Therefore, from (3.10) and (3.9), it can be verified that $(Y_{k,m})_{PCT}$ gets a small value instead of being thresholded to zero. This is also expected since a PWP coefficient that has a less probability to be a noise coefficient should not be thresholded to zero.

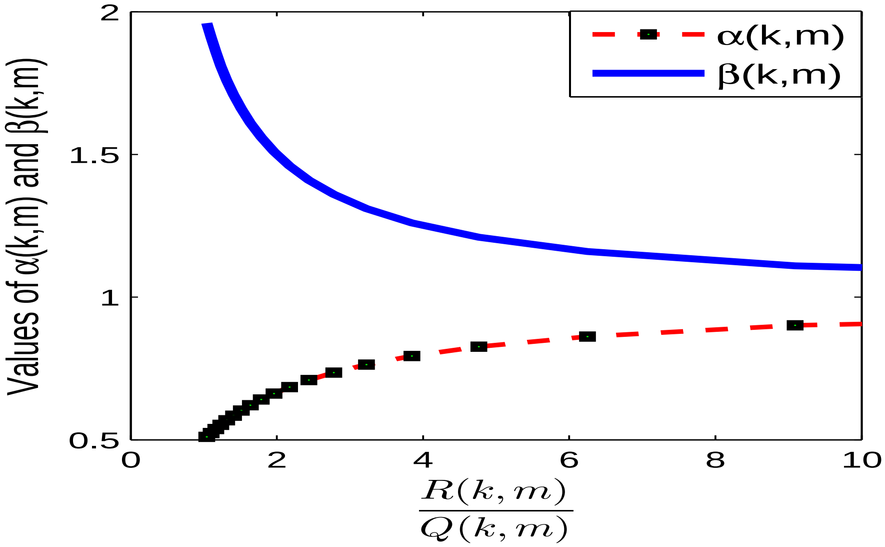


Fig. 3.13: Plot of $\alpha(k,m)$ and $\beta(k,m)$ for different values of $\frac{R(k,m)}{q(k,m)}$

Determination of Shape Parameters

The proposed thresholding function can be adapted to noise characteristics of the input noisy speech based on the shape parameters $\alpha(k,m)$ and $\beta(k,m)$ which are defined as

$$\alpha(k,m) = \frac{1 + R(k,m)}{2(1 + Q(k,m))}, \tag{3.13}$$

$$\beta(k,m) = \frac{2(1+Q(k,m))}{(1+R(k,m))}, \tag{3.14}$$

where $R(k,m)$ and $Q(k,m)$ are the speech presence and absence probabilities, respectively, of the m-th coefficient in the k-th subband.

Given two hypotheses,H_0 and H_1, which indicate respectively speech absence and presence in the m-th coefficient of the k-th subband, and assuming a complex Gaussian distributions for both speech and noise PWP coefficients [6], the conditional pdfs of the speech and noise PWP coefficients are given by

$$f(Y(k,m)|H_0(k,m)) = \frac{1}{\pi\sigma_n^2} exp(-\frac{|Y(k,m)|^2}{\sigma_n^2}), \tag{3.15}$$

$$f(Y(k,m)|H_1(k,m)) = \frac{1}{\pi(\sigma_n^2+\sigma_r^2)} exp(-\frac{|Y(k,m)|^2}{\sigma_n^2+\sigma_r^2}). \tag{3.16}$$

Using *aposteriori* and *apriori* SNRs defined by [6]

$$\Upsilon(k,m) = \frac{|Y(k,m)|^2}{\sigma_n^2(k,m)}, \tag{3.17}$$

$$\eta(k,m) = \frac{\sigma_r^2(k,m)}{\sigma_n^2(k,m)}, \tag{3.18}$$

and following (7.13) and (7.14), the conditional pdfs of the *aposteriori* SNR can be written as [35]

$$f(\Upsilon(k,m)|H_0(k,m)) = e^{-\Upsilon(k,m)} I_2, \tag{3.19}$$

$$f(\Upsilon(k,m)|H_1(k,m)) = \frac{1}{1+\eta(k,m)} \times exp(-\frac{\Upsilon(k,m)}{1+\eta(k,m)}) I_2. \tag{3.20}$$

In (7.17) and (7.18), $I_2 = u(\Upsilon(k,m))$ is the unit step function. Noting that the conditional speech presence probability $R(k,m) = P(H_1(k,m)|\Upsilon(k,m))$, applying Bayes rule and using (7.18), an expression for $R(k,m)$ can be derived as

$$R(k,m) = [1 + \frac{Q(k,m)}{1-Q(k,m)}(1+\widehat{\eta}(k,m)) exp(-v(k,m))]^{-1}, \tag{3.21}$$

where $\widehat{\eta}(k,m)$ is the estimated *apriori* SNR obtained as in [35] and

$$v(k,m) = \frac{\widehat{\eta}(k,m)\Upsilon(k,m)}{(1+\widehat{\eta}(k,m))}. \tag{3.22}$$

Speech absence probability $Q(k,m)$ in (7.19) can be determined as

$$Q(k,m) = 1 - R_{local}(k,m)R_{global}(k,m)R_{subband}(k,m). \tag{3.23}$$

In (7.21), $R_{local}(k,m)$ and $R_{global}(k,m)$ are the speech presence probabilities in local and global windows in the PWP domain. Letting τ for representing either "local" or "global" window, $R_\tau(k,m)$ can be given by

$$R_\tau(k,m) = \begin{cases} 0, & \text{if } \xi_\tau(k,m) \leq \xi_{min} \\ 1, & \xi_\tau(k,m) \geq \xi_{max}, \\ \frac{log(\xi_\tau(k,m)/\xi_{min})}{log(\xi_{max}/\xi_{min})}, & \text{otherwise}, \end{cases} \tag{3.24}$$

where $\xi_\tau(k,m)$ representing either "local" or "global" average of the *apriori* SNR given by

$$\xi_\tau(k,m) = \sum_{i=-W_\tau}^{i=W_\tau} h_\tau(i)\xi(k-i,m). \tag{3.25}$$

In (7.23), h_τ is a normalized window of size $2w_\tau+1$ and $\xi(k,m)$ represents a recursive average of the *apriori* SNR given by

$$\xi(k,m) = \kappa\xi(k,m-1) + (1-\kappa)\widehat{\eta}(k,m-1), \tag{3.26}$$

where κ denotes a smoothing constant. Note that in (7.22), ξ_{min} and ξ_{max} are the two empirical constants representing minimum and maximum values of $\xi(k,m)$ given in (7.24). $R_{subband}(k)$ in (7.21) can be computed as

$$R_{subband}(k) = \begin{cases} 0, & \text{if } \xi_{subband}(k) < \xi_{min} \\ 1, & \text{if } \xi_{subband}(k) > \xi_{subband}(k-1) and \xi_{subband}(k) > \xi_{min}, \\ \mu(k), & \text{otherwise}, \end{cases} \tag{3.27}$$

where $\mu(k)$ is expressed as

$$\mu(k) = \begin{cases} 0, & \text{if } \xi_{subband}(k) \leq \xi_{peak}(k)\xi_{min} \\ 1, & \text{if } \xi_{subband}(k) \geq \xi_{peak}(k)\xi_{max}, \\ \frac{log(\xi_{subband}(k)/\xi_{peak}(k)/\xi_{min})}{log(\xi_{max}/\xi_{min})}, & \text{otherwise}. \end{cases} \tag{3.28}$$

In (7.26) and (7.25), $\xi_{subband}(k)$ is determined as

$$\xi_{subband}(k) = \frac{1}{N_c} \sum_{1 \ll m \ll N_c} \xi(k,m) \tag{3.29}$$

and ξ_{peak} in (7.25) is a confined peak value of $\xi_{subband}(k)$. Thus computing $R(k,m)$ and $Q(k,m)$ following (7.19) and (7.21), the shape parameters $\alpha(k,m)$ and $\beta(k,m)$ can be determined using (7.11) and (7.12), respectively.

3.1.6 Inverse Perceptual Wavelet Packet Transform

For a noisy speech frame, we obtain thresholded PWP coefficients using the proposed threshold in (3.7) and the proposed thresholding function in (3.9). An enhanced speech frame $\widehat{r}[n]$ is synthesized by performing inverse PWP transform as

$\widehat{r}[n] = PWP^{-1}(Y_{k,m})_{PCT}$.

The enhanced speech signal is reconstructed by using the standard overlap-and-add method [18].

3.2 Results Considering Gaussian Statistical Model

In this Section, a number of simulations is carried out to evaluate the performance of the proposed method considering Gaussian statistical model.

3.2.1 Simulation Conditions

Real speech sentences from the NOIZEUS database are employed for the experiments, where the speech data is sampled at 8 KHz [42]. To imitate a noisy environment, noise sequence is added to the clean speech samples at different SNR levels ranging from 15 dB to -15 dB. As in [43], two different types of noises, such as car and babble are adopted from the NOIZEUS databases [42].

In order to obtain overlapping analysis frames, hamming windowing operation is performed, where the size of each of the frame is 512 samples with 50% overlap between successive frames. A 6-level PWP decomposition tree with 10 db bases function is applied on the noisy speech frames [38], [40] resulting in subbands $k = 1, 2, ...24$.

The values of used constants to determine the shape parameters in the proposed thresholding function are given in table 7.1.

Table 3.2: Constants used to determine the shape parameters

Constants	Value
β	0.7
ξ_{min}	-10 dB
ξ_{max}	-5 dB
ξ_{peak}	10 dB
w_{local}	1
w_{global}	15

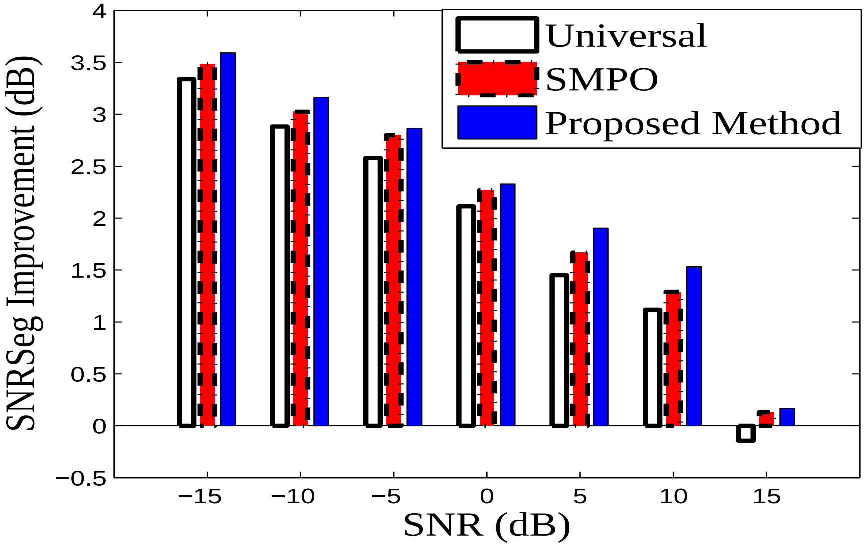


Fig. 3.14: SNRSeg Improvement for different methods in car noise

3.2.2 Comparison Metrics

Standard Objective metrics namely, Segmental SNR (SNRSeg) improvement in dB, Perceptual Evaluation of Speech Quality (PESQ) and Weighted Spectral Slope (WSS) are used for the evaluation of the proposed method [1]. The proposed method is subjectively evaluated in terms of the spectrogram representations of the clean speech, noisy speech and enhanced speech. Formal listening tests are also carried out in order to find the analogy between the objective metrics and the subjective sound quality. The performance of our method is compared with some of the state-of-the-art speech enhancement methods, such as Universal [10] and SMPO [43] in both objective and subjective senses.

Table 3.3: PESQ for different methods in car noise

SNR(dB)	Universal	SMPO	Proposed Method
-15	1.16	1.15	1.27
-10	1.23	1.37	1.45
-5	1.32	1.51	1.61
0	1.43	1.69	1.79
5	1.69	2.07	2.13
10	1.93	2.38	2.43
15	2.14	2.60	2.75

3.2.3 Objective Evaluation

Results for Speech signals with Car Noise

SNRSeg improvement, PESQ and WSS for speech signals corrupted with car noise for Universal, SMPO and proposed methods are shown in Fig.7.10, Table 3.3 and Fig.7.11.

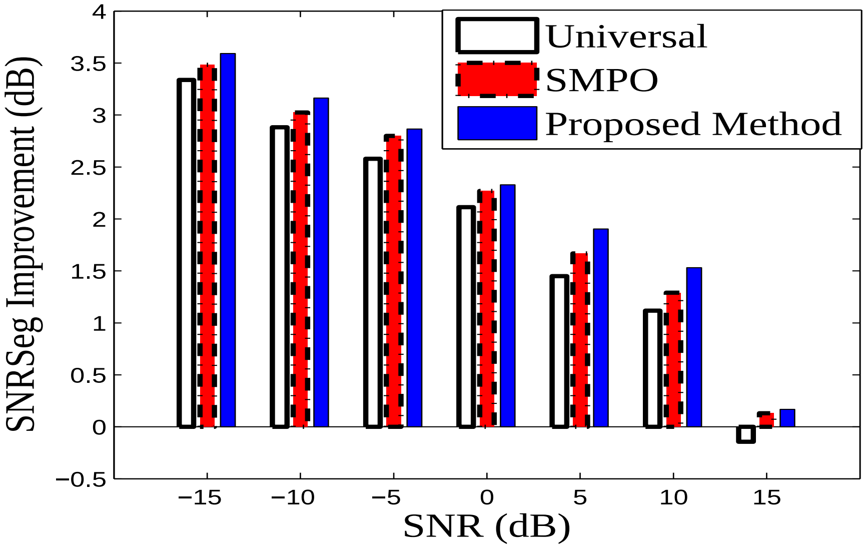


Fig. 3.15: SNRSeg Improvement for different methods in car noise

In Fig.7.10, the performance of the proposed method is compared with that of the other methods at different levels of SNR for car noise in terms of Segmental SNR improvement. We see, the SNRSeg improvement increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields the highest SNRSeg improvement. Such larger values of SNRSeg improvement at a low level of SNR attest the capability of the proposed method in producing enhanced speech with better quality for speech severely corrupted by car noise.

In Table 3.3, it can be seen that at a low level of SNR, such as $-15dB$, all

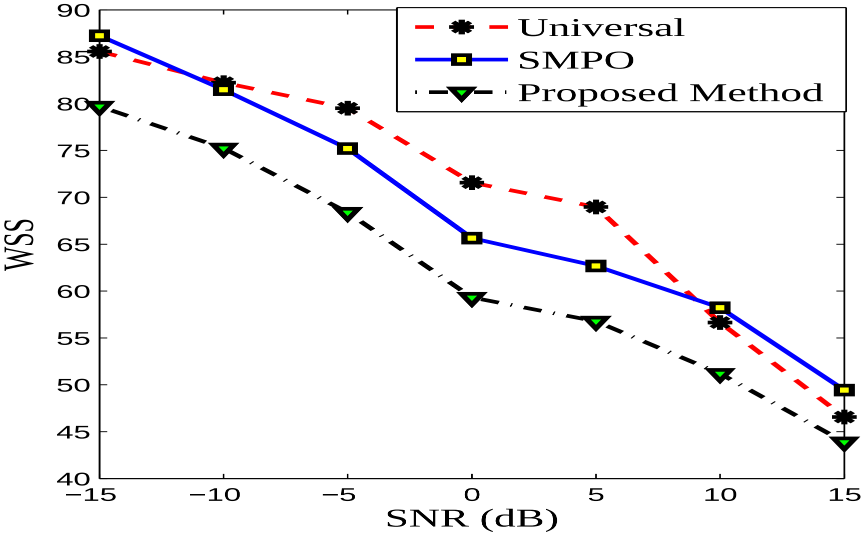


Fig. 3.16: WSS for different methods in car Noise

the methods show lower values of PESQ scores, whereas the PESQ score is much higher, as expected, for the proposed method. The proposed method also yields larger PESQ scores compared to that of the other methods at higher levels of SNR. Since, at a particular SNR, a higher PESQ score indicates a better speech quality, the proposed method is indeed better in performance in the presence of a car noise.

Fig.7.11 represents the WSS values as a function of SNR for the proposed method and that for the other methods. As shown in the figure, the WSS values resulting from all other methods are relatively larger for a wide range of SNR levels, whereas the proposed method is capable of producing enhanced speech with better quality as it gives lower values of WSS even at a low SNR of $-15dB$.

Results for Speech signals with Multi-talker Babble Noise

SNRSeg improvement, PESQ and WSS for speech signals corrupted with babble noise for Universal, SMPO and proposed methods are shown in Fig.7.12, 7.14 and 7.13, respectively.

In Fig. 7.12, it can be seen that at a low level of SNR of $-15dB$, the proposed method provides a SNRSeg improvement that is significantly higher than that of the methods of comparison. The proposed method still shows better performance in terms of SNRSeg improvement for higher SNRs also.

For speech corrupted with babble noise, in Fig.7.13, the mean values of PESQ

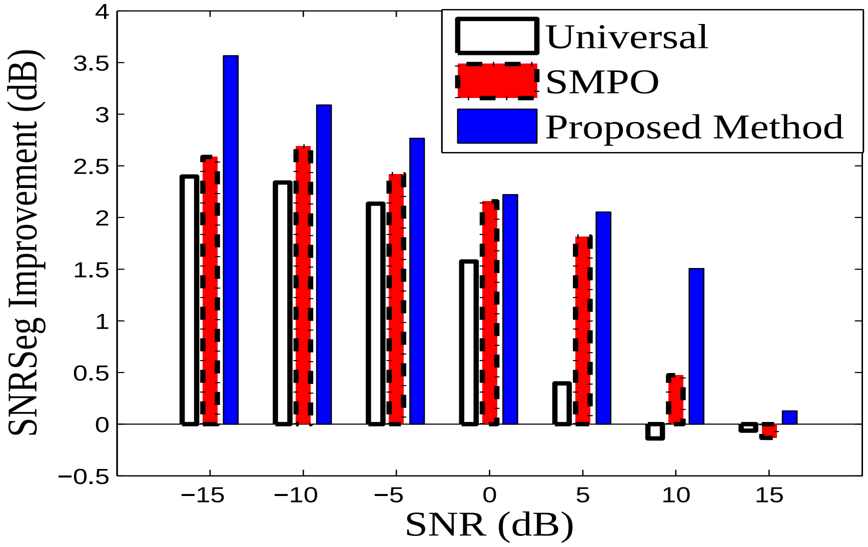


Fig. 3.17: SNRSeg Improvement for different methods in babble noise

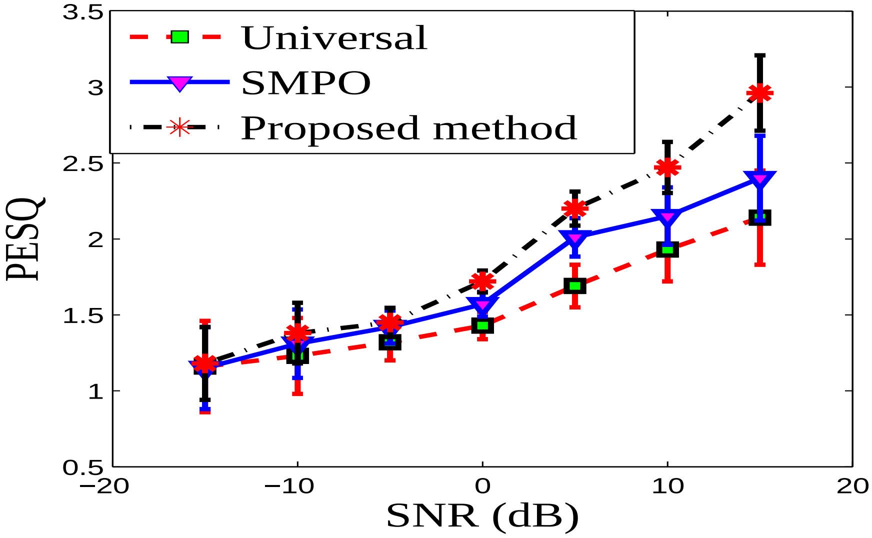


Fig. 3.18: PESQ for different methods in babble noise

with standard deviation obtained using the proposed method is plotted and compared with that of the other methods. From this plot, it is seen that over the whole SNR range considered, the proposed method continue to provide higher PESQ with almost non-overlapping standard deviation in the presence of babble noise.

The performance of the proposed method is compared with that of the other methods in terms of WSS in Fig.7.14 at different levels of SNRs in presence of babble noise. It is clearly seen from this figure that WSS increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields a WSS that is significantly lower

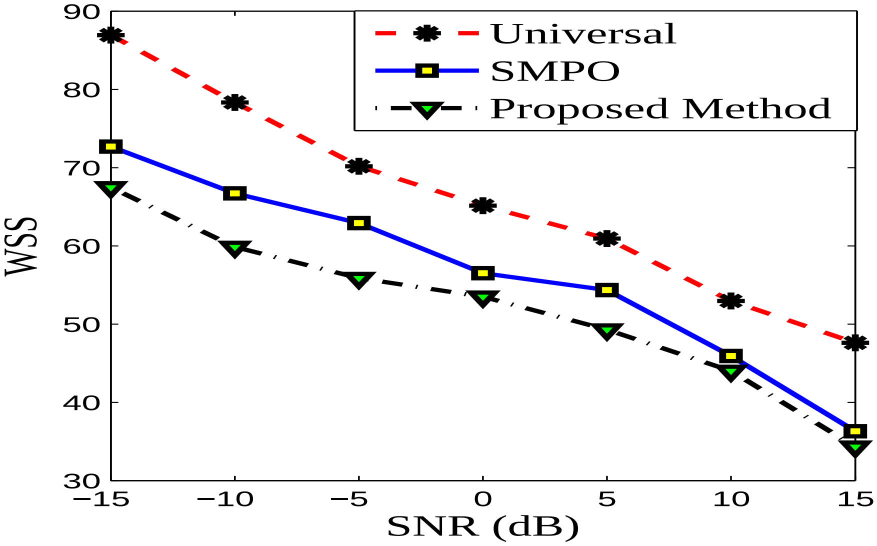


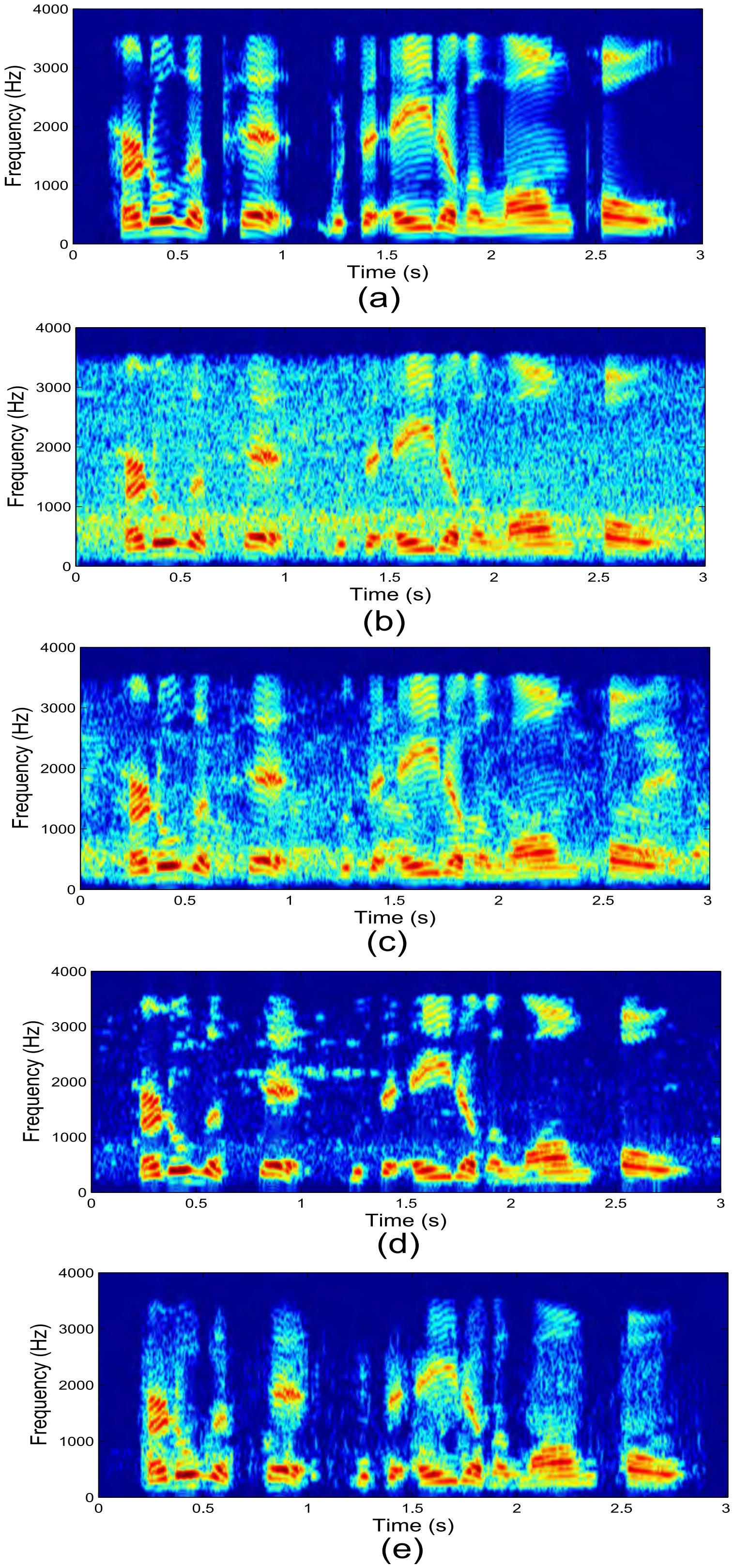
Fig. 3.19: WSS for different methods in babble noise

than that of all other methods, which remains lower over the higher SNRs also.

3.2.4 Subjective Evaluation

In order to evaluate the subjective observation of the enhanced speech, spectrograms of the clean speech, the noisy speech, and the enhanced speech signals obtained by using the proposed method and all other methods are presented in Fig. 7.15 for car noise corrupted speech at an SNR of 10 dB. It is evident from this figure that the harmonics are well preserved and the amount of distortion is greatly reduced in the proposed method. Thus, the spectrogram observations with lower distortion also validate our claim of better speech quality as obtained in our objective evaluations in terms of higher SNR improvement in dB, higher PESQ score and lower WSS in comparison to the other methods. Another set of spectrograms for babble noise corrupted speech at an SNR of 10 dB is also presented in Fig.7.16. This figure attests that the proposed method has a better efficacy in preserving speech harmonics even in case of babble noise.

Formal listening tests are also conducted, where ten listeners are allowed and arranged to perceptually evaluate the enhanced speech signals. A full set (thirty sentences) of the NOIZEUS corpus was processed by Universal, SMPO and proposed method for subjective evaluation at different SNRs. Subjective tests were performed according to ITU-T recommendation P.835 [42]. In this tests, a listener



(a)

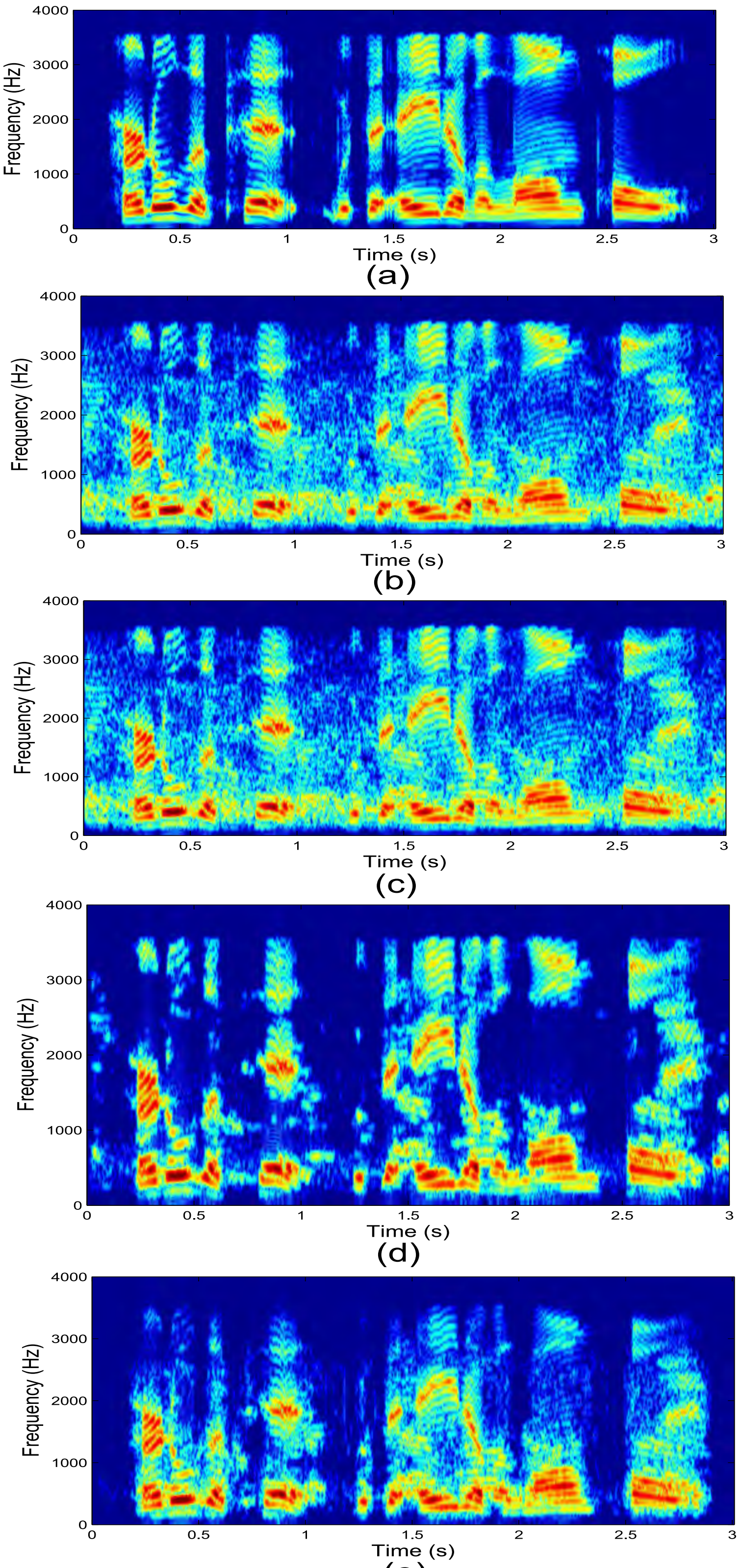
(b)

(c)

(d)

(e)

Fig. 3.20: Spectrograms of (a) Clean Signal (b) Noisy Signal with 10dB car noise; spectrograms of enhanced speech from (c) Universal method (d) SMPO method (e) Proposed Method



(a)

(b)

(c)

(d)

(e)

Fig. 3.21: Spectrograms of (a) Clean Signal (b) Noisy Signal with 10dB babble noise; spectrograms of enhanced speech from (c) Universal method (d) SMPO method (e) Proposed Method

Table 3.4: Mean Score of SIG scale for different methods in presence of car noise at 5 db

Listener	Universal	SMPO	Proposed Method
1	3.6	4.0	4.0
2	3.3	3.9	3.7
3	3.9	4.0	4.2
4	3.4	4.2	4.5
5	3.2	3.8	4.0
6	2.9	3.6	3.9
7	3.8	3.8	4.2
8	3.5	3.7	4.2
9	3.5	3.9	3.8
10	3.7	3.9	4.0

is instructed to successively attend and rate the enhanced speech signal based on (a) the speech signal alone using a scale of SIG (1 = very unnatural, 5 = very natural), (b) the background noise alone using a scale of background conspicuous/ intrusiveness (BAK) (1 = very conspicuous, very intrusive; 5 = not noticeable), and (c) the overall effect using the scale of the mean opinion score (OVRL) (1 = bad, 5 = excellent). More details about the testing methodology can be found in [44]. The mean scores of SIG, BAK, and OVRL scales for the three speech enhancement methods evaluated in the presence of car noise at an SNR of 5 dB are shown in Tables 3.4, 3.5, and 3.6. For the three methods evaluated using babble noise-corrupted speech at an SNR of 10 dB, the mean scores of SIG, BAK, and OVRL scales are also summarized in Tables 3.7, 3.8, and 3.9. The mean scores in the presence of both car and babble noises demonstrate that the lower signal distortion (i.e., higher SIG scores) and the lower noise distortion (i.e., higher BAK scores) are obtained with the proposed method relative to that obtained by Universal and SMPO methods in most of the conditions. It is also shown that a consistently better performance in OVRL scale is offered by the proposed method not only in car but also in babble noisy conditions at both SNR levels considered in comparison to that provided by all the methods mentioned above. Overall, it is found that the proposed method possesses the highest subjective sound quality in comparison to that of the other methods in case of different noises at various levels of SNRs.

Table 3.5: Mean Score of BAK scale for different methods in presence of car noise at 5 db

Listener	**Universal**	**SMPO**	**Proposed Method**
1	4.0	4.5	5.0
2	4.3	4.9	4.7
3	4.2	4.4	4.9
4	4.4	4.7	4.8
5	4.2	4.8	4.7
6	3.9	4.6	4.9
7	3.8	3.9	4.4
8	4.4	4.6	4.6
9	3.5	3.8	4.5
10	4.2	4.5	4.8

Table 3.6: Mean Score of OVL scale for different methods in presence of car noise at 5 db

Listener	**Universal**	**SMPO**	**Proposed Method**
1	2.6	4.0	4.1
2	3.3	3.8	3.7
3	3.9	4.1	4.3
4	3.6	4.2	4.2
5	3.3	3.9	4.1
6	3.9	4.6	4.9
7	3.8	3.8	4.3
8	3.6	4.1	4.2
9	3.5	4.5	4.7
10	3.9	4.6	4.8

Table 3.7: Mean Score of SIG scale for different methods in presence of Babble noise at 5 db

Listener	**Universal**	**SMPO**	**Proposed Method**
1	3.6	4.0	4.0
2	3.3	3.9	3.7
3	4.2	3.9	4.0
4	3.4	4.2	4.5
5	3.2	3.8	4.0
6	2.9	3.6	3.9
7	3.8	3.8	4.2
8	3.4	3.6	4.1
9	3.5	3.9	3.7
10	3.7	3.8	3.9

Table 3.8: Mean Score of BAK scale for different methods in presence of Babble noise at 5 db

Listener	**Universal**	**SMPO**	**Proposed Method**
1	4.0	4.5	5.0
2	4.3	4.9	4.7
3	4.2	4.4	4.9
4	4.4	4.7	4.8
5	4.2	4.8	4.7
6	3.9	4.6	4.9
7	3.8	3.9	4.4
8	4.4	4.6	4.7
9	3.5	3.9	4.7
10	4.7	4.8	4.9

Table 3.9: Mean Score of OVL scale for different methods in presence of babble noise at 5 db

Listener	**Universal**	**SMPO**	**Proposed Method**
1	2.6	4.0	4.1
2	3.3	3.8	3.7
3	3.9	4.1	4.3
4	3.6	4.2	4.2
5	3.3	3.9	4.1
6	3.9	4.6	4.9
7	3.8	3.8	4.3
8	3.6	4.1	4.2
9	3.5	4.5	4.7
10	3.9	4.8	4.9

3.3 Conclusion

In this paper, we developed a Gaussian statistical model-based technique for the TE operated PWP coefficients of the noisy speech in order to obtain a suitable threshold value. By employing the proposed gaussian pdf dependent custom thresholding function, the PWP coefficients of the noisy speech are thresholded in order to obtain an enhanced speech. Simulation results show that the proposed method yields consistently better results in the sense of higher Segmental SNR Improvement in dB, higher output PESQ, and lower WSS values than those of the existing methods. The improved performance of the proposed method is also indicated and attested by the much better spectrogram outputs and in terms of the higher scores in the formal subjective listening tests.

Chapter 4

Speech Enhancement Using Laplace Modeling of Teager Energy Operated Perceptual Wavelet Packet Coefficients

In this chapter, speech enhancement based on Laplace modeling of TE operated PWP coefficients is described [45]. An adaptive threshold is determined analytically using the Laplace model of TE operated PWP coefficients and then this threshold is imposed upon the PWP coefficients of noisy speech using custom thresholding function which is devised as a combination of μ-law and semisoft thresholding functions. Detail simulation is performed to compare the proposed method with the state-of-the art speech enhancement techniques.

4.0.1 Proposed Laplace Distribution Model for TE Operated PWP Coefficients

Following discussion in chapter 3, As an alternative to formulate a pdf of the $t_{k,m}^{j}$ of speech, we can easily formulate the histogram of the $t_{k,m}^{j}$ and approximate the histogram by a reasonably close probability distribution function, namely Laplace distribution in place of Gaussian distribution [40]. For the $t_{k,m}$s in a subband of a noisy speech frame, the empirical histogram along with the Gaussian and the Laplace distributions are superimposed in Fig. 4.1, 4.2 and 4.3 in presence of car noise at SNRs of -15, 0 and 15 dB. From this figure, it is obvious that Laplace distribution fits the empirical histogram better than the Gaussian distribution. Similar analysis results are obtained for empirical histogram, Gaussian and Laplace distribution of TE operated noise PWP coefficients at the same SNRs as used in Fig. 4.1, 4.2

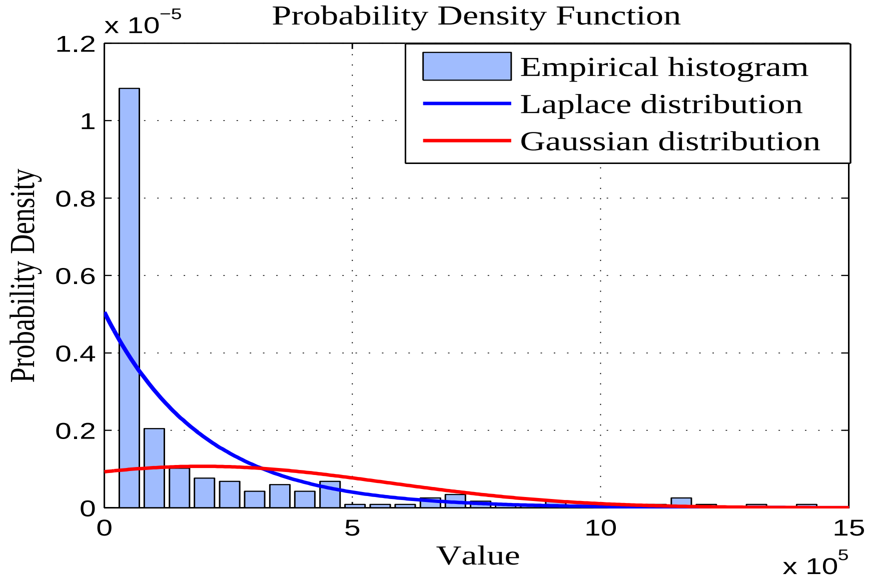


Fig. 4.1: Empirical histogram, Gaussian and Laplace distributions of TE operated PWP coefficients of noisy speech at SNR of -15 dB

and 4.3 and are shown in Fig. 4.4, 4.5 and 4.6. Such statistical matching between the Gaussian and Laplace is also explained in terms of AIC index [46]. It can be noted from [46] that the more negative value of the AIC index indicates the more matching between two pdfs. Assuming Gaussian and Laplace distributions for $t_{k,m}$ in a subband of a noisy speech frame, mean values of AIC index obtained using different speech sentences are shown in Fig. 7.7 for a range of SNR -15dB to 15dB in the presence of car noise. From Fig. 7.7, it is clearly attested that the Laplace distribution offers better matching with the empirical histogram compared to the Gaussian distribution not only at SNR of 15dB but also at an SNR as low as -15dB. The plot representing the values of AIC index for the Gaussian and Laplace distributions of $t_{k,m}$ of noise at SNR level ranging from -15dB to 15dB is illustrated in Fig. 7.8. This figure shows that AIC index for $t_{k,m}$ of noise continues to exhibit more negative values for Laplace distribution thus maintaining better pdf matching for a wide range of SNR. Therefore, we propose to approximate the histograms of $t_{k,m}$ of noisy speech, noise and clean speech by Laplace distribution and perform statistical modeling for calculating the threshold adaptive to different subbands.

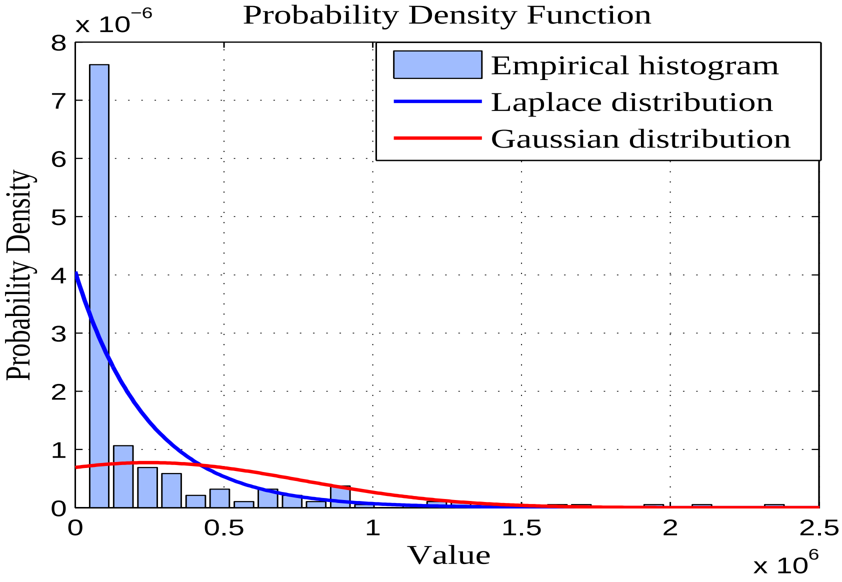


Fig. 4.2: Empirical histogram, Gaussian and Laplace distributions of TE operated PWP coefficients of noisy speech at SNR of 0 dB

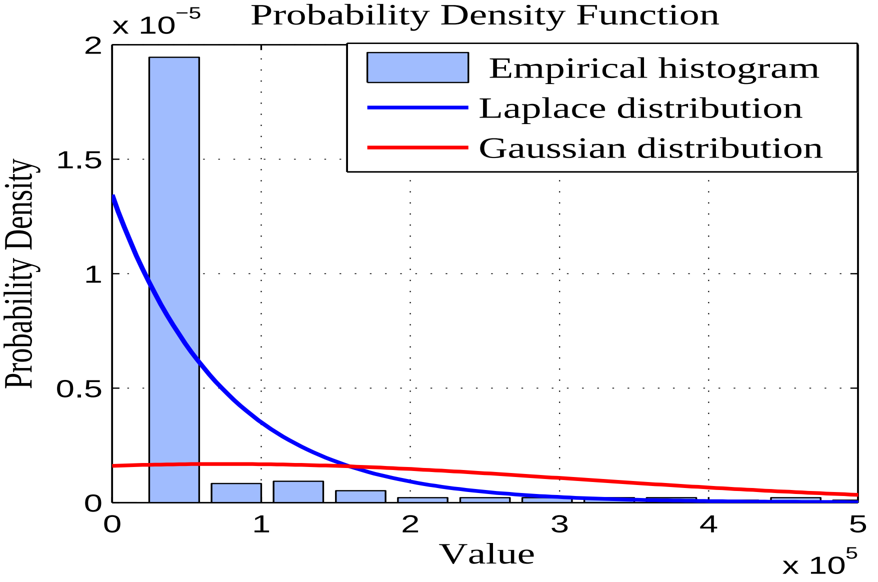


Fig. 4.3: Empirical histogram, Gaussian and Laplace distributions of TE operated PWP coefficients of noisy speech at SNR of 15 dB

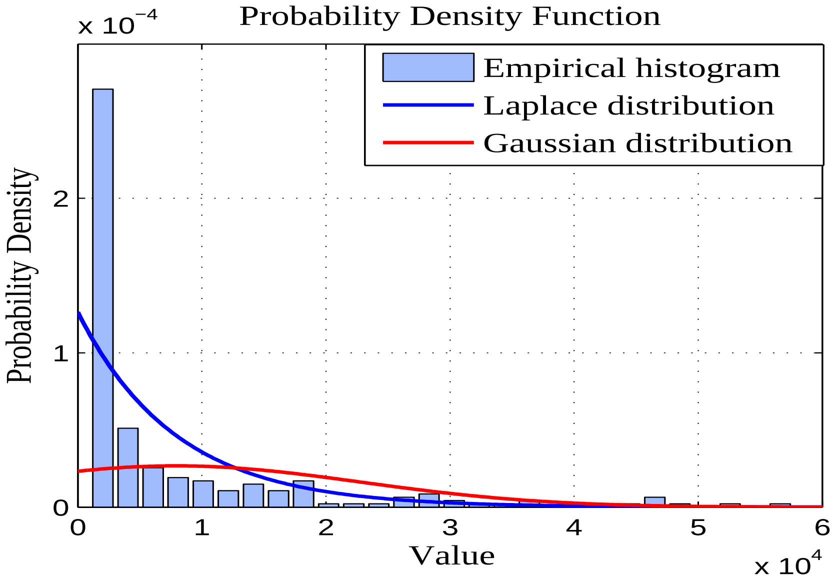


Fig. 4.4: Empirical histogram, Gaussian and Laplace distributions of TE operated PWP coefficients of noise at SNR of -15 dB

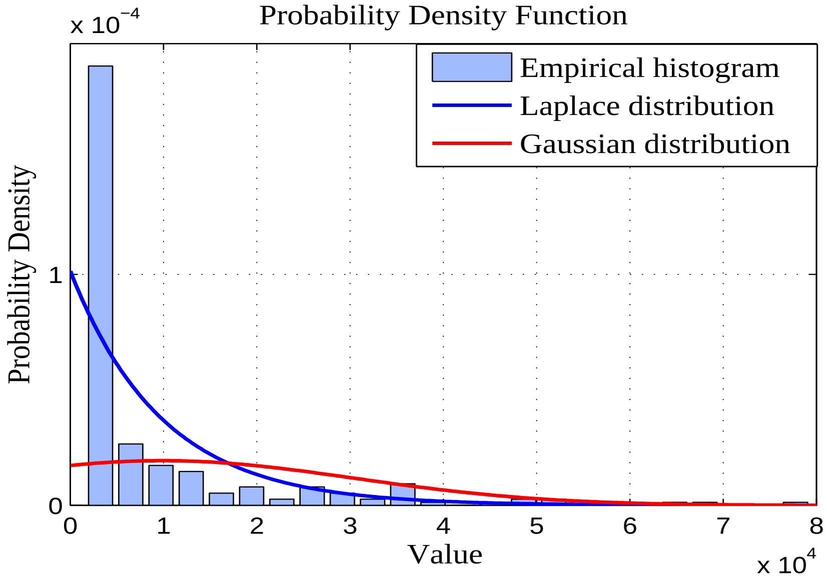


Fig. 4.5: Empirical histogram, Gaussian and Laplace distributions of TE operated PWP coefficients of noisy speech at SNR of 0 dB

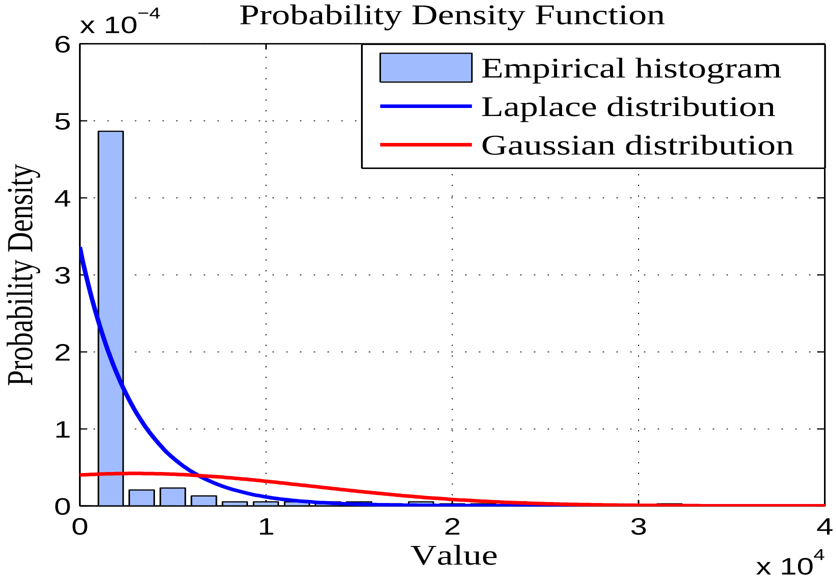


Fig. 4.6: Empirical histogram, Gaussian and Laplace distributions of TE operated PWP coefficients of noisy speech at SNR of 15 dB

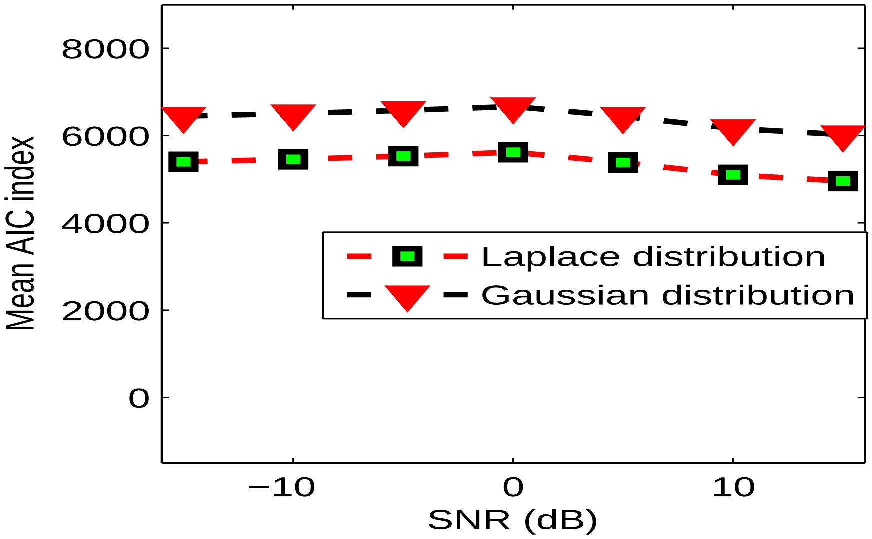


Fig. 4.7: Mean values of AIC index of TE operated PWP coefficients of (a) noisy speech (b) noise assuming Gaussian and Laplace distributions

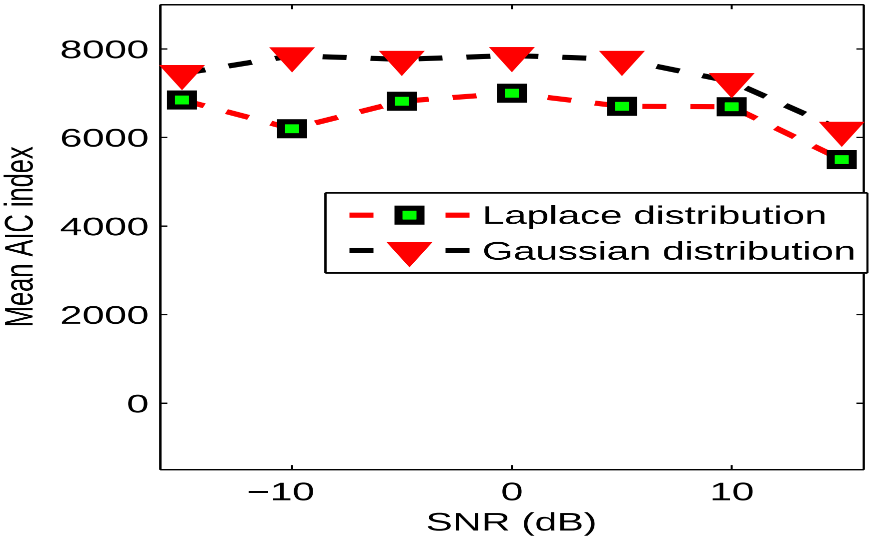


Fig. 4.8: Mean values of AIC index of TE operated PWP coefficients of (a) noisy speech (b) noise assuming Gaussian and Laplace distributions

4.0.2 Proposed Adaptive Threshold Calculation assuming Laplace distribution

Following the discussion in chapter 3, laplace distribution pdf for $p_i(t_{k,m})$ can be written as

$$p_i(t_{k,m}) = \frac{v}{2\sigma_s^2} \times exp(-\frac{|x|}{\sigma_s^2}) \tag{4.1}$$

Where σ_s^2 represents the power of $t_{k,m}$ of noisy speech.

Letting σ_r^2 as the power of $t_{k,m}$ of clean speech and σ_n^2 as the power of $t_{k,m}$ of noise and using the fact $\sigma_s^2 = \sigma_r^2 + \sigma_n^2$, we can write

$$p_i(t_{k,m}) = \frac{v}{2\sqrt{\sigma_r^2 + \sigma_n^2}} \times exp(-\frac{|x|}{\sqrt{\sigma_r^2 + \sigma_n^2}}) \tag{4.2}$$

Following (4.1) in a similar way, laplace pdf for $q_i(t_{k,m})$ can also be written as

$$q_i(t_{k,m}) = \frac{v}{2\sigma_n^2} \times exp(-\frac{|x|}{\sigma_n^2}) \tag{4.3}$$

By substituting (4.2) and (4.3) in (3.6), we obtain

$$\int_1^{\lambda} [\frac{v}{2\sigma_s^2} \times exp(-\frac{|x|}{\sigma_s^2}) - \frac{1}{2\sigma_n^2} \times exp(-\frac{|x|}{\sigma_n^2})] I_1 dx = 0 \tag{4.4}$$

where I_1 is defined as, $I_1 = ln(1-v) \times exp(-\frac{|x|}{\sigma_s^2} + \frac{|x|}{\sigma_n^2})$

By solving (4.4), value of λ can be derived as

$$\lambda(k) = \sqrt{\frac{\sigma_n^2(k)(1+\gamma(k)) \times log(1+\gamma(k))}{\gamma(k)}} \tag{4.5}$$

where $\gamma(k)$ is the segmental SNR of subband k defined as

$$\gamma(k) = \frac{\sigma_r^2(k)}{\sigma_n^2(k)}. \tag{4.6}$$

The proposed threshold $\lambda(k)$ in (4.5) derived assuming Laplace pdf is compared with that obtained assuming Gaussian pdf given by

$$\lambda(k) = \frac{\sigma_n(k)}{\sqrt{\gamma_k}}\sqrt{2(\gamma_k + \gamma_k^2)} \times ln(\sqrt{1+\frac{1}{\gamma_k}}) \tag{4.7}$$

in Fig.4.9. This figure shows that the pattern of the threshold value is similar for both the pdfs at high as well as low SNRs. In terms of value, although Laplace pdf shows slightly lower values at high SNRs, but the threshold values are much lower than that of the Gaussian pdf specially at low SNRs. Therefore, the threshold derived from the Laplace pdf offers less chance of removing speech coefficients while performing thresholding operation not only at high SNR but also at difficult low SNRs.

The proposed threshold as derived in (4.5) is high for higher noise power and low for lower noise power thus is adaptive to noise power of different subbands. In this method, voice activity detector is not needed as the threshold is automatically adapted to the silent and speech frames. At a silent frame, since noise power is significantly higher than the signal power, the proposed threshold results in a higher value as seen from (4.5). Such a value imposes more coefficients to be thresholded thus removing noise coefficients completely at subbands of a silent frame. Note that, in this paper, noise is estimated using Improved Minima Controlled Recursive Averaging (IMCRA) method [35].

4.0.3 Proposed Thresholding Function

We propose a custom thresholding function derived from the meu law and the semisoft thresholding functions [41]. Representing $\lambda(k)$ derived from (4.5) as $\lambda_1(k)$ and letting $\lambda_2(k) = 2\lambda_1(k)$, the proposed thresholding function is developed as

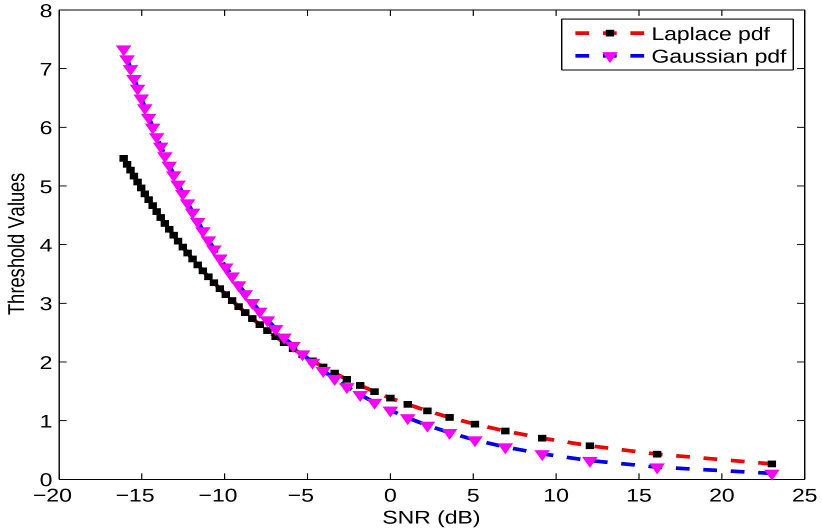


Fig. 4.9: Comparison of threshold values with respect to SNR for Laplace and Gaussian pdfs

$$(Y_{k,m})_{PCT} = \begin{cases} \alpha\frac{sgn(Y^j_{k,m}).|Y^j_{k,m}|}{\mu}[(1+\mu)^{\frac{|Y^j_{k,m}|}{\lambda_1(k)}} - 1], & \text{if } |(Y^j_{k,m})| \leq \lambda_1 \\ Y_{k,m}, & \text{if } |(Y^j_{k,m})| \geq \lambda_2(k), \\ (1-\alpha)Z_1 + \alpha Z_2, & \text{otherwise} \end{cases} \tag{4.8}$$

where

$$Z_1 = sgn(Y^j_{k,m}) \times \lambda_2(k)\frac{|(Y^j_{k,m})| - \lambda_1(k)}{\lambda_2(k) - \lambda_1(k)} \tag{4.9}$$

$$Z_2 = Y^j_{k,m} \tag{4.10}$$

In (4.8), $(Y_{k,m})_{PCT}$ stands for the PWP coefficients thresholded by the proposed custom thresholding function expressed and shape parameter of the proposed thresholding function is represented by $\alpha(k, m)$.

The comparison of the proposed custom thresholding function with the conventional meu law and semisoft thresholding functions is shown in Fig. 4.10. In the region between λ_1 and λ_2, this figure demonstrates the flexibility of the proposed thresholding operation in a sense that it can be viewed as $(1 - \alpha(k, m))(Y_{k,m})_{SS} + \alpha(k, m)(Y_{k,m})_{ML}$ which is a linear combination of the meu law and the semisoft thresholding function. Here, $(Y_{k,m})_{ML}$ stands for the PWP coefficients thresholded by the meu law thresholding function and $(Y_{k,m})_{SS}$ represents the PWP coefficients thresholded by the semisoft thresholding function. Unlike these functions, depending on the value of shape parameter $\alpha(k, m)$, it can be verified from (3.9) that the proposed thresholding function gets the following forms,

$$\lim_{\alpha(k,m)\to 0} (Y_{k,m})_{PCT} = (Y_{k,m})_{SS}$$
$$\lim_{\alpha(k,m)\to 1} (Y_{k,m})_{PCT} = (Y_{k,m})_{ML}$$

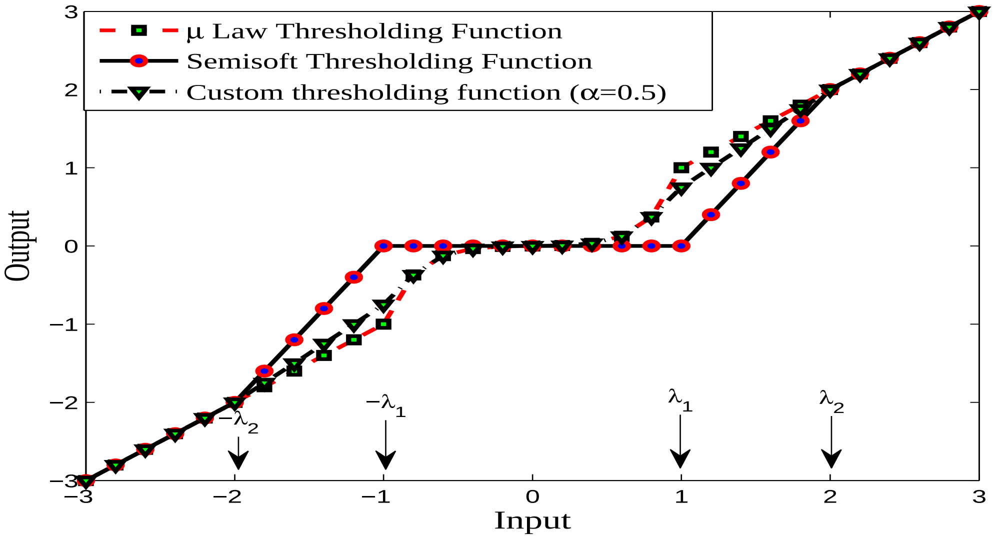


Fig. 4.10: Input Output Relation for semisoft, μ law and proposed custom thresholding function

The enhanced speech frame is synthesized by performing the inverse PWP transformation PWP^{-1} on the resulting thresholded PWP coefficients and The final enhanced speech signal is reconstructed by using the standard overlap-and-add method.

4.1 Results Considering Laplace Statistical Model

In this Section, a number of simulations is carried out with the same simulation conditions as described in chapter 3 to evaluate the performance of the proposed method considering Laplace statistical model. Same comparison metrics are used to compare the proposed method with the previously mentioned comparison methods.

Table 4.1: PESQ for different methods in Car Noise assuming Laplace distribution for the proposed method

SNR(dB)	SMPO	Universal	Proposed Method (Laplace)
-15	1.26	1.13	1.27
-10	1.43	1.27	1.45
-5	1.61	1.41	1.61
0	1.77	1.53	1.79
5	2.12	1.77	2.13
10	2.43	1.97	2.43
15	2.66	2.15	2.75

4.1.1 Objective Evaluation

Results for Speech signals with Car Noise

SNRSeg, WSS and PESQ scores for speech signals corrupted with car noise for Universal, SMPO and proposed method are shown in Fig.4.11, 4.12 and Table. 4.1.

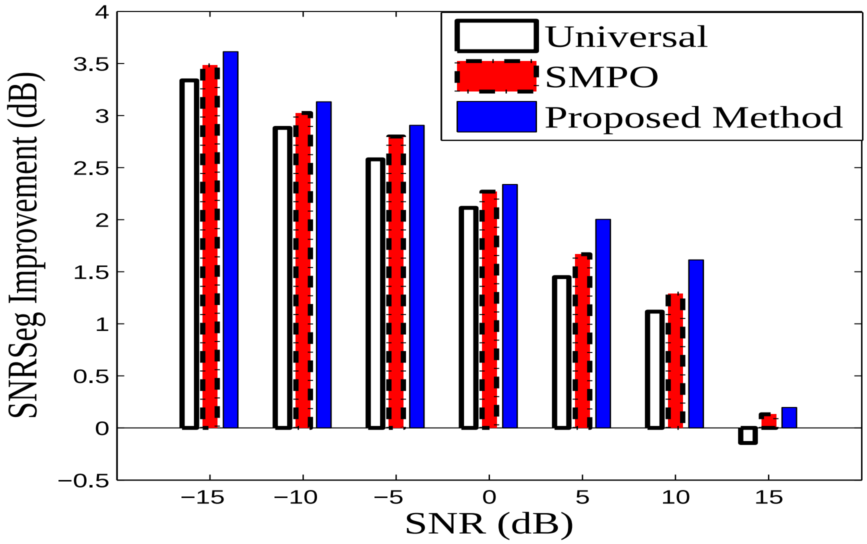


Fig. 4.11: SNRSeg Improvement for different methods in Car Noise assuming Laplace distribution for the proposed method

In Figure 4.11, the performance of the proposed method is compared with that of the other methods at different levels of SNR for car noise in terms of Segmental SNR inprovement. We see, the SNRSeg improvement increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields the highest SNRSeg improvement. Such larger values of SNRSeg improvement at a low level of SNR attest the capability of the proposed method in producing enhanced speech with better quality for speech severely corrupted by car noise.

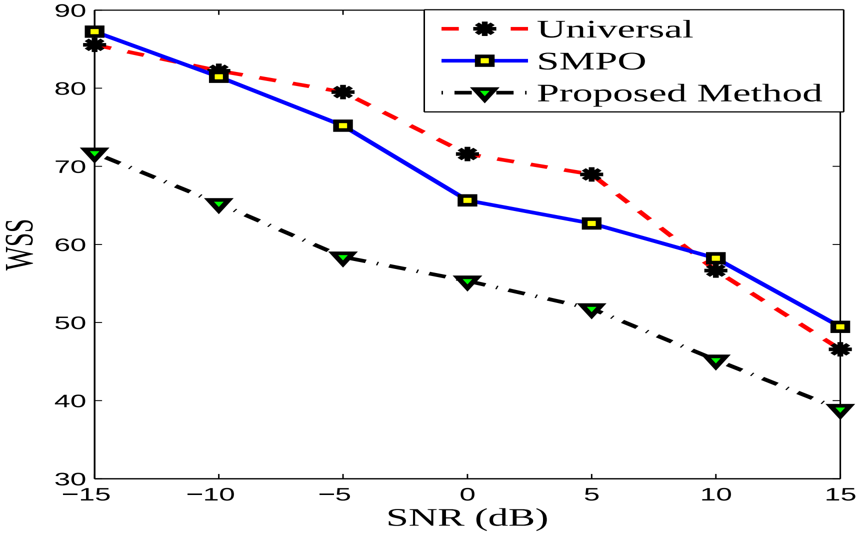


Fig. 4.12: WSS for different methods in Car Noise assuming Laplace distribution for the proposed method

In Table.4.1, it can be seen that at a low level of SNR, such as $-15dB$, all the methods show lower values of PESQ scores, whereas the PESQ score is much higher, as expected, for the proposed method. The proposed method also yields larger PESQ scores compared to that of the other methods at higher levels of SNR. Since, at a particular SNR, a higher PESQ score indicates a better speech quality, the proposed method is indeed better in performance in the presence of a car noise.

Fig.4.12 represents the WSS values as a function of SNR for the proposed method and that for the other methods. As shown in the figure, the WSS values resulting from all other methods are relatively larger for a wide range of SNR levels, whereas the proposed method is capable of producing enhanced speech with better quality as it gives lower values of WSS even at a low SNR of $-15dB$.

Results for Speech signals with Multi-talker Babble Noise

SNRSeg improvement, PESQ and WSS for speech signals corrupted with babble noise for Universal, SMPO and proposed methods are shown in Fig.4.13, 4.15 and 4.14, respectively.

In Fig. 4.13, it can be seen that at a low level of SNR of $-15dB$, the proposed method provides a SNRSeg improvement that is significantly higher than that of the methods of comparison. The proposed method still shows better performance in terms of SNRSeg improvement for higher SNRs also.

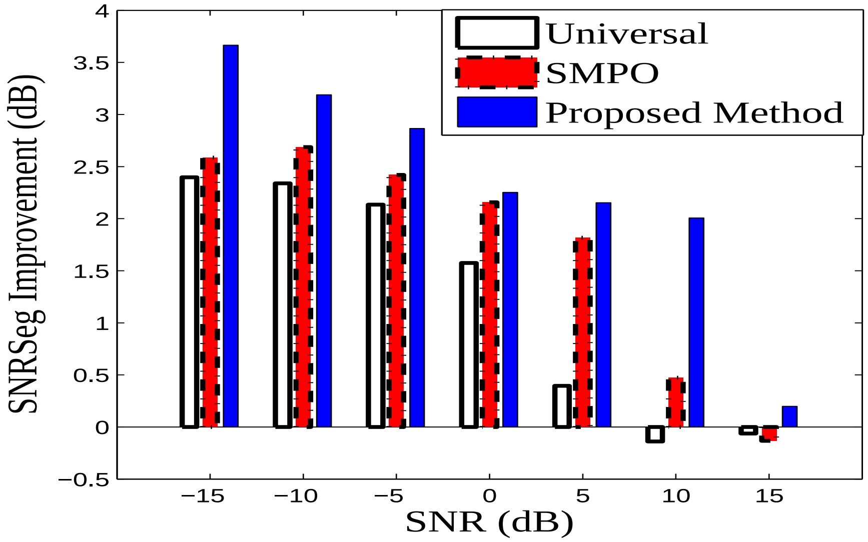


Fig. 4.13: SNRSeg Improvement for different methods in Babble Noise assuming Laplace distribution for the proposed method

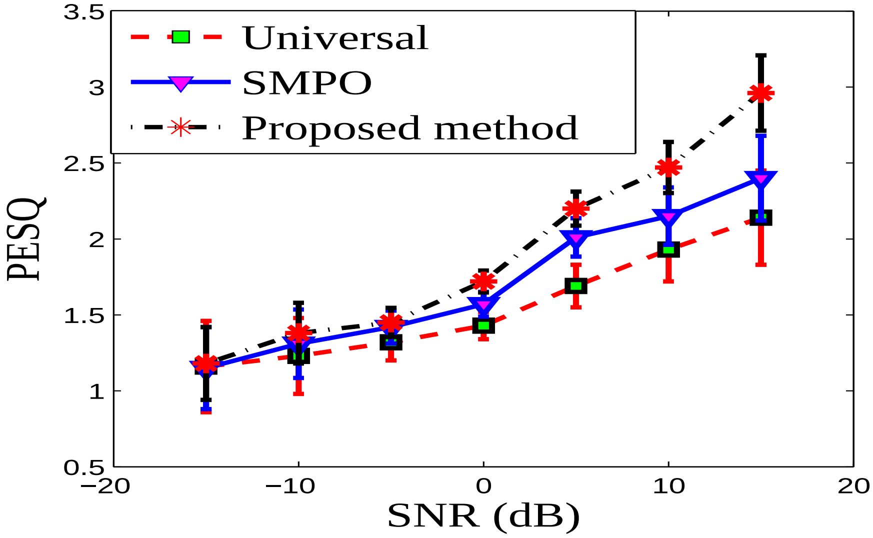


Fig. 4.14: PESQ for different methods in Babble Noise assuming Laplace distribution for the proposed method

For speech corrupted with babble noise, in Fig.4.14, the mean values of PESQ with standard deviation obtained using the proposed method is plotted and compared with that of the other methods. From this plot, it is seen that over the whole SNR range considered, the proposed method continue to provide higher PESQ with almost non-overlapping standard deviation in the presence of babble noise.

The performance of the proposed method is compared with that of the other methods in terms of WSS in Fig.4.15 at different levels of SNRs in presence of babble

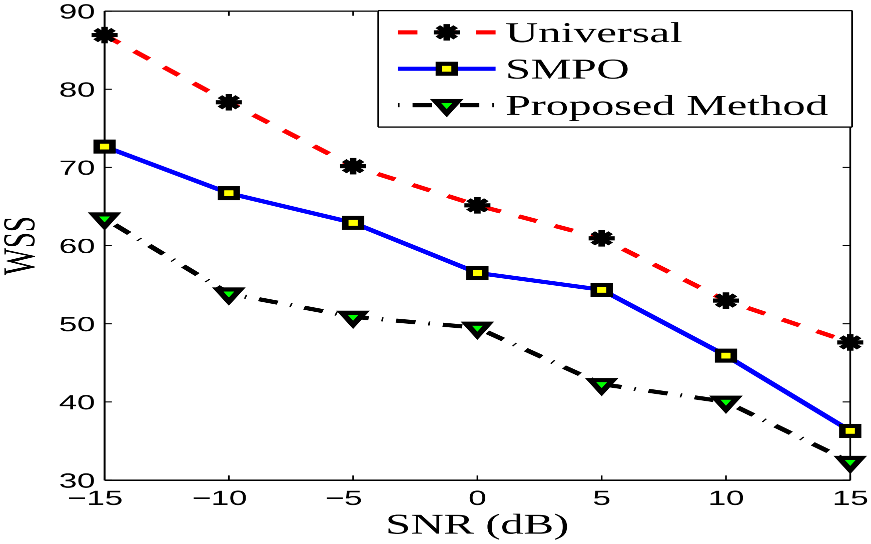


Fig. 4.15: WSS for Babble Noise for different methods in Babble Noise assuming Laplace distribution for the proposed method

noise. It is clearly seen from this figure that WSS increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields a WSS that is significantly lower than that of all other methods, which remains lower over the higher SNRs also.

4.1.2 Subjective Evaluation

In order to evaluate the subjective observation of the enhanced speech, spectrograms of the clean speech, the noisy speech, and the enhanced speech signals obtained by using the proposed method and all other methods are presented in Fig. 4.16 for car noise corrupted speech at an SNR of 10 dB. It is evident from this figure that the harmonics are well preserved and the amount of distortion is greatly reduced in the proposed method. Thus, the spectrogram observations with lower distortion also validate our claim of better speech quality as obtained in our objective evaluations in terms of higher SNR improvement in dB, higher PESQ score and lower WSS in comparison to the other methods. Another set of spectrograms for babble noise corrupted speech at an SNR of 10 dB is also presented in Fig.4.17. This figure attests that the proposed method has a better efficacy in preserving speech harmonics even in case of babble noise.

The mean scores of SIG, BAK, and OVRL scales for the three speech enhancement methods evaluated in the presence of car noise at an SNR of 5 dB are shown in Tables 4.2, 4.3, and 4.4. For the three methods examined using babble noise-

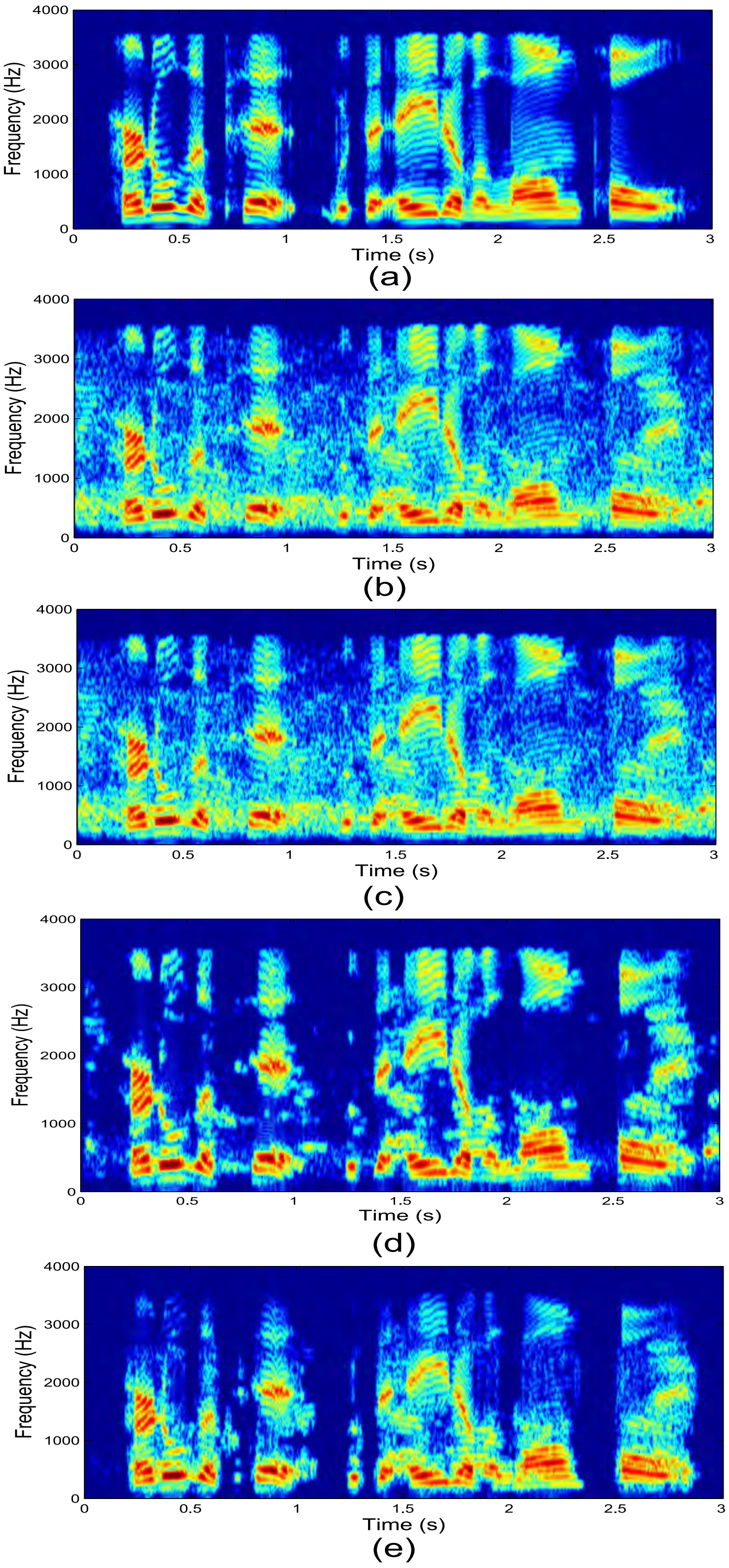


Fig. 4.16: Spectogram of Output for noisy signal mixed with 10dB car noise for different methods (a) Clean Signal (b) Noisy Signal (c) SMPO (d) Universal (e) Proposed Method assuming Laplace distribution

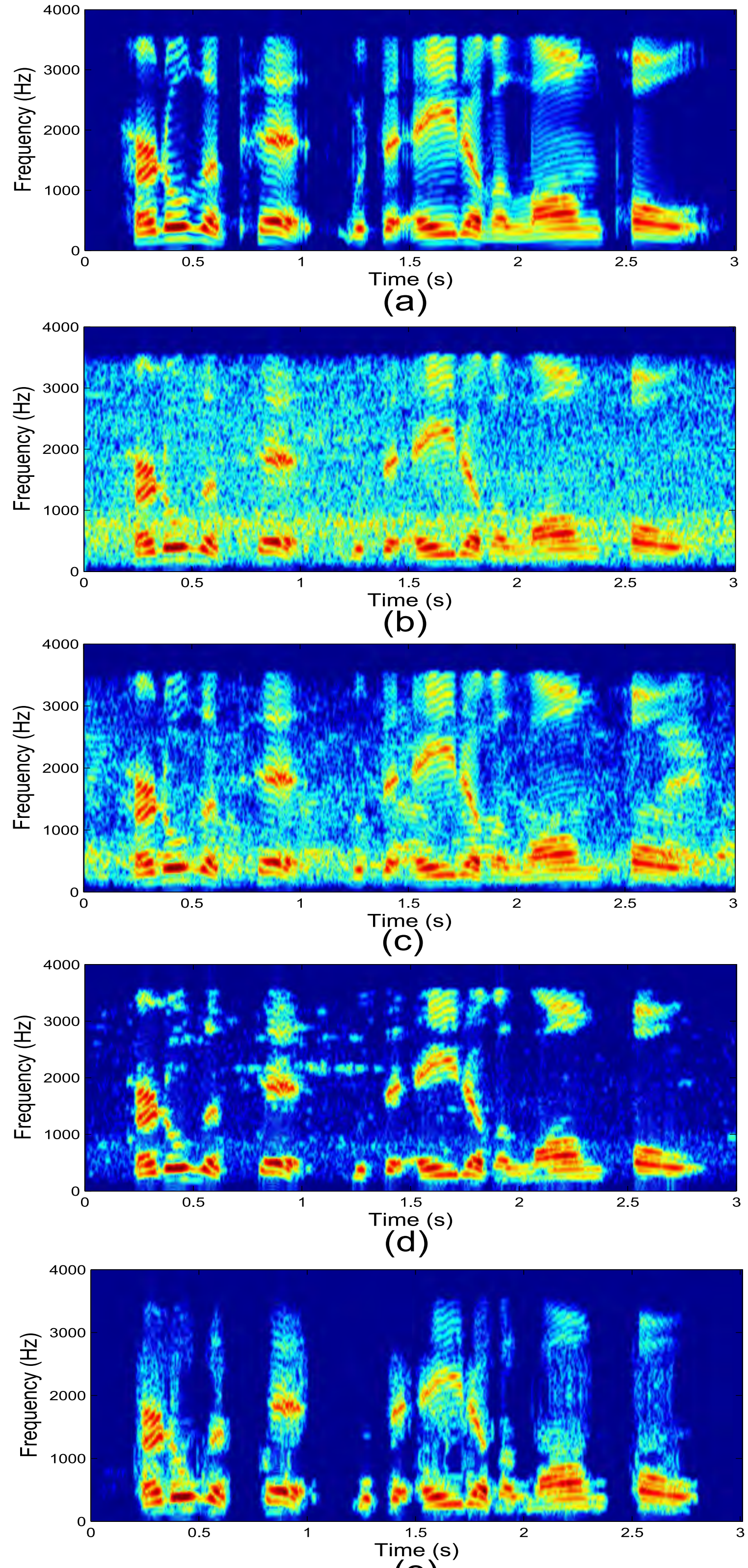


Fig. 4.17: Spectogram of Output for noisy signal mixed with 10dB babble noise for different methods (a) Clean Signal (b) Noisy Signal (c) SMPO (d) Universal (e) Proposed Method assuming Laplace distribution

Table 4.2: Mean Score of SIG scale for different methods in presence of car noise at 5 db assuming Laplace distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Laplace)**
1	4.0	3.6	4.0
2	3.9	3.3	3.7
3	4.0	3.9	4.2
4	4.2	3.4	4.5
5	3.8	3.2	4.0
6	3.6	2.9	3.9
7	3.8	3.8	4.2
8	3.7	3.5	4.2
9	3.9	3.5	3.8
10	3.9	3.7	4.0

Table 4.3: Mean Score of BAK scale for different methods in presence of car noise at 5 db assuming Laplace distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Laplace)**
1	4.5	4.0	5.0
2	4.9	4.3	4.7
3	4.4	4.2	4.9
4	4.7	4.4	4.8
5	4.8	4.2	4.7
6	4.6	3.9	4.9
7	3.9	3.8	4.4
8	4.6	4.4	4.6
9	3.8	3.5	4.5
10	4.5	4.2	4.8

corrupted speech at an SNR of 10 dB, the mean scores of SIG, BAK, and OVRL scales are summarized in Tables 4.5, 4.6, and 4.7. The mean scores in the presence of both car and babble noises demonstrate that the lower signal distortion (i.e., higher SIG scores) and the lower noise distortion (i.e., higher BAK scores) are obtained with the proposed method relative to that obtained by Universal and SMPO methods in most of the conditions. It is also shown that a consistently better performance in OVRL scale is offered by the proposed method not only in car but also in all other noisy conditions at both SNR levels of considered in comparison to that provided by all the methods mentioned above. Overall, it is found that the proposed method possesses the highest subjective sound quality in comparison to that of the other methods in case of different noises at various levels of SNR.

Table 4.4: Mean Score of OVL scale for different methods in presence of car noise at 5 db assuming Laplace distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Laplace)**
1	4.0	2.6	4.1
2	3.8	3.3	3.7
3	4.1	3.9	4.3
4	4.2	3.6	4.2
5	3.9	3.3	4.1
6	4.6	3.9	4.9
7	3.8	3.8	4.3
8	4.1	3.6	4.2
9	4.5	3.5	4.7
10	4.6	3.9	4.8

Table 4.5: Mean Score of SIG scale for different methods in presence of Babble noise at 5 db assuming Laplace distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Laplace)**
1	4.0	3.6	4.0
2	3.9	3.3	3.7
3	4.0	3.9	4.2
4	4.2	3.4	4.5
5	3.8	3.2	4.0
6	3.6	2.9	3.9
7	3.8	3.8	4.2
8	3.6	3.4	4.1
9	3.9	3.5	3.7
10	3.8	3.7	3.9

Table 4.6: Mean Score of BAK scale for different methods in presence of Babble noise at 5 db assuming Laplace distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Laplace)**
1	4.0	2.6	4.1
2	3.8	3.3	3.7
3	4.1	3.9	4.3
4	4.2	3.6	4.2
5	3.9	3.3	4.1
6	4.6	3.9	4.9
7	3.8	3.8	4.3
8	4.1	3.6	4.2
9	4.5	3.5	4.7
10	4.8	3.9	4.9

Table 4.7: Mean Score of OVL scale for different methods in presence of Babble noise at 5 db assuming Laplace distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Laplace)**
1	4.0	2.6	4.1
2	3.8	3.3	3.7
3	4.1	3.9	4.3
4	4.2	3.6	4.2
5	3.9	3.3	4.1
6	4.6	3.9	4.9
7	3.8	3.8	4.3
8	4.1	3.6	4.2
9	4.5	3.5	4.7
10	4.8	3.9	4.9

4.2 Conclusion

To solve the problems of speech enhancement, an improved perceptual wavelet packet based approach using the Laplace pdf of Teager Energy Operated wavelet Packet coefficients has been presented in this paper. We incorporated a statistical model-based technique with teager energy operator of the wavelet packet coefficients to obtain a suitable threshold using symmetric K-L divergence. For solving the equation of pdf's, we choose Laplace distribution as an acceptable pdf for noisy speech, clean speech and noise TE operated PWP coefficients in each sub-band. Unlike the unique threshold based method, the threshold value here is adapted based on the speech and silence segments. Then, by employing the proposed custom thresholding function, the PWP coefficients of the noisy speech are thresholded in order to obtain a cleaner speech. Simulation results show that the proposed method yields consistently better results in the sense of higher output SNR in dB, higher output PESQ, and lower WSS values than those of the existing methods, hence results in a better enhanced speech.

Chapter 5

Speech Enhancement Using Rayleigh Modeling of Teager Energy Operated Perceptual Wavelet Packet Coefficients

In this chapter, speech enhancement based on Rayleigh modeling of TE operated PWP coefficients is described [47]. An adaptive threshold is determined analytically using the Rayleigh model of TE operated PWP coefficients and then this threshold is imposed upon the PWP coefficients of noisy speech using custom thresholding function which is devised as a combination of modified hard and semisoft thresholding functions. Detail simulation has been performed to compare the proposed method with the state-of-the art speech enhancement techniques.

5.0.1 Proposed Rayleigh Distribution Model for TE Operated PWP Coefficients

Following discussion in chapter 3, As an alternative to formulate a pdf of the of speech, we can easily formulate the histogram of its $t_{k,m}$ and can approximate the histogram by a reasonably close pdf namely Gaussian and Rayleigh distribution for$_1$0.$Forthet_{k,m}$s in a subband of a noisy speech frame, the empirical histogram along with the Gaussian and the Rayleigh distributions are superimposed in Fig. 5.1, 5.2 and 5.3 in presence of car noise at SNRs of -15, 0 and 15 dB. From this figures, it is obvious that Rayleigh distribution fits the empirical histogram better than the Gaussian distribution. Similar analysis results are obtained for empirical histogram, Gaussian and Rayleigh distribution of TE operated noise PWP coefficients at the same SNRs as used in Fig. 5.1, 5.2 and 5.3 and are shown in Fig. 5.4, 5.5 and

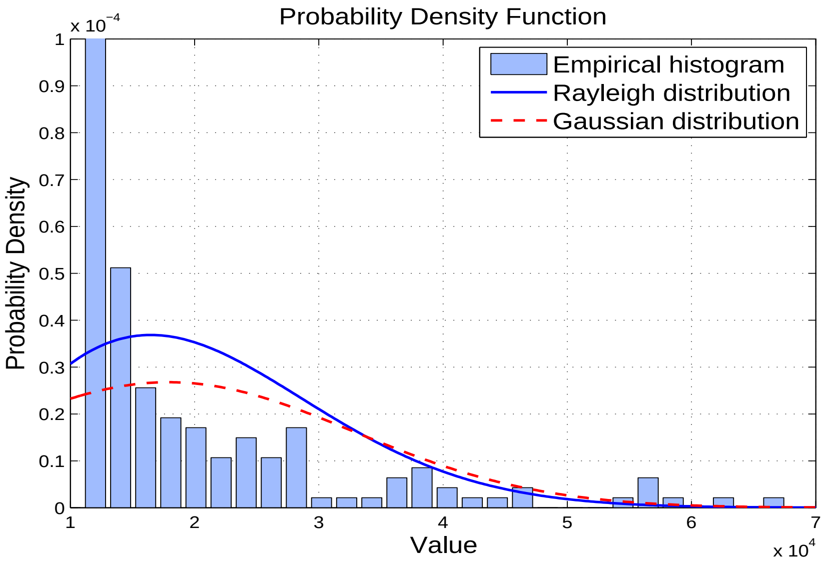


Fig. 5.1: Empirical histogram, Gaussian and Rayleigh distributions of TE operated PWP coefficients of noisy speech at SNRs of -15 dB

5.6. Such statistical matching between the Gaussian and Rayleigh distribution is also explained in terms of AIC index [46]. It can be noted from [46] that the more negative value of the AIC index indicates the more matching between two pdfs. Assuming Gaussian and Rayleigh distributions for $t_{k,m}$ in a subband of a noisy speech frame, mean values of AIC index obtained using different speech sentences are shown in Fig. 7.7 for a range of SNR -15dB to 15dB in the presence of car noise. From Fig. 7.7, it is clearly attested that the Rayleigh distribution offers better matching with the empirical histogram compared to the Gaussian distribution not only at SNR of 15dB but also at an SNR as low as -15dB. The plot representing the values of AIC index for the Gaussian and Rayleigh distributions of $t_{k,m}$ of noise at SNR level ranging from -15dB to 15dB is illustrated in Fig. 7.8. This figure shows that AIC index for $t_{k,m}$ of noise continues to exhibit more negative values for Rayleigh distribution thus maintaining better pdf matching for a wide range of SNR. Therefore, we propose to approximate the histograms of $t_{k,m}$ of noisy speech, noise and clean speech by Rayleigh distribution and perform statistical modeling for calculating the threshold adaptive to different subbands.

5.0.2 Proposed Adaptive Threshold Calculation assuming Rayleigh distribution

Following the discussion in chapter 3, Rayleigh pdf for $p_i(t_{k,m})$ can be written as

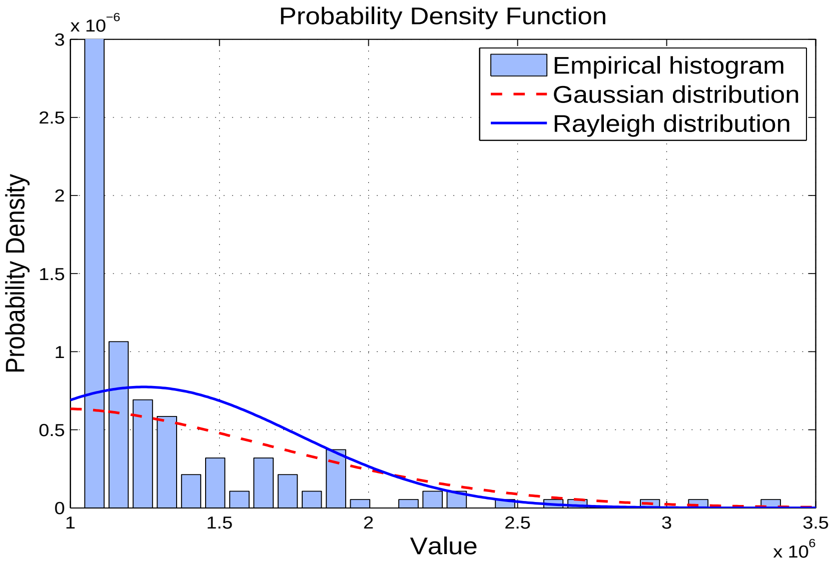


Fig. 5.2: Empirical histogram, Gaussian and Rayleigh distributions of TE operated PWP coefficients of noisy speech at SNRs of 0 dB

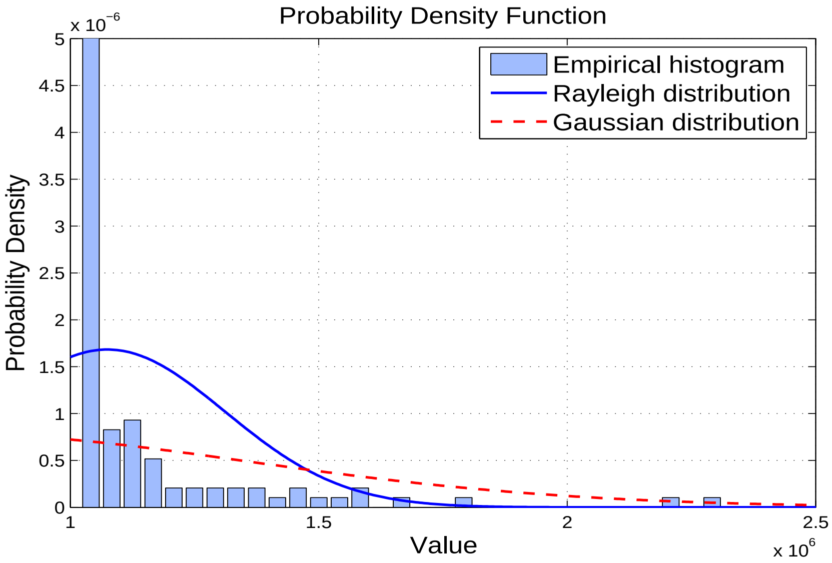


Fig. 5.3: Empirical histogram, Gaussian and Rayleigh distributions of TE operated PWP coefficients of noisy speech at SNRs of 15 dB

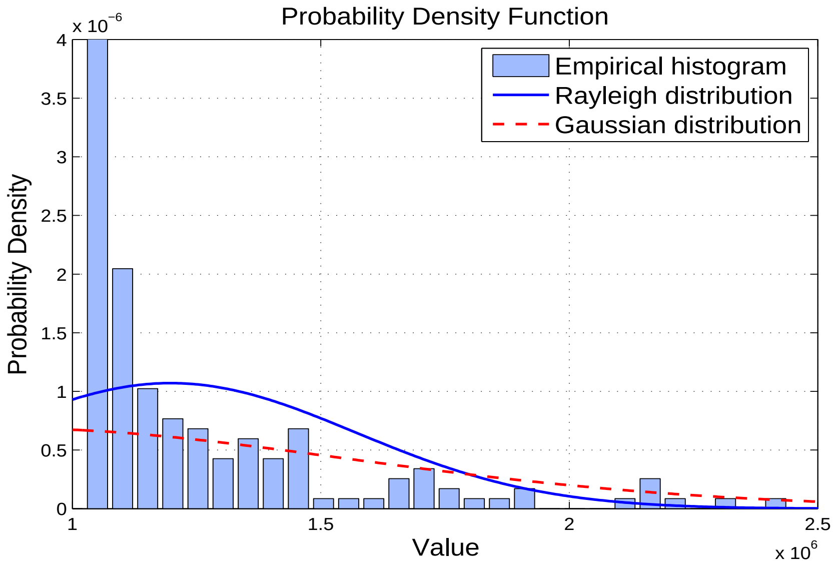


Fig. 5.4: Empirical histogram, Gaussian and Rayleigh distributions of TE operated noise PWP coefficients at an SNR of -15 dB

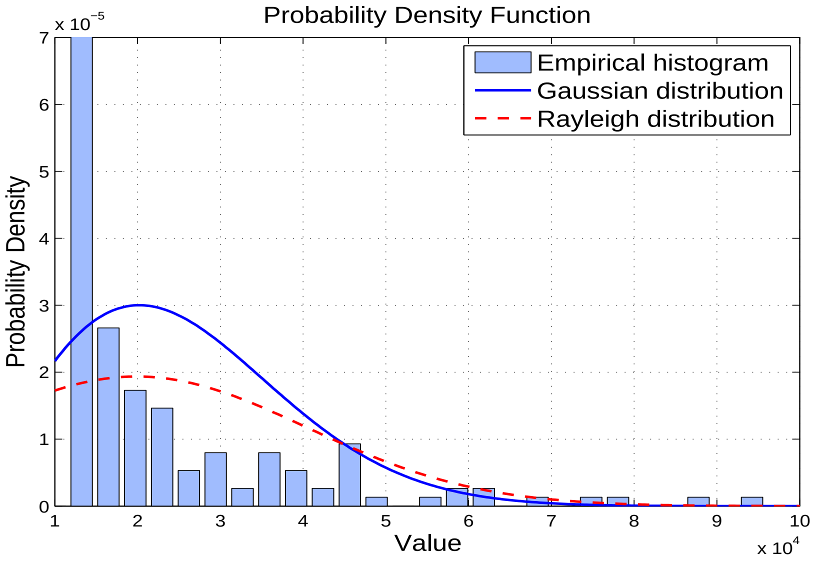


Fig. 5.5: Empirical histogram, Gaussian and Rayleigh distributions of TE operated noise PWP coefficients at an SNR of 0 dB

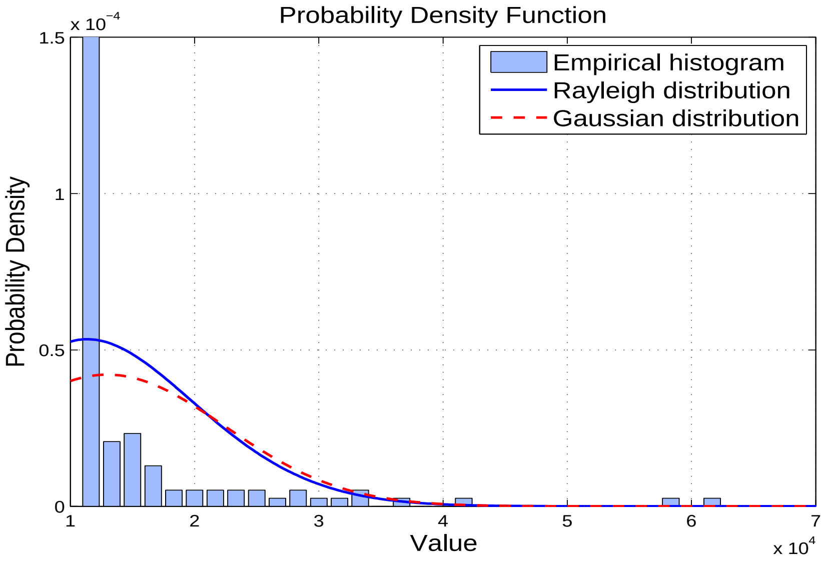


Fig. 5.6: Empirical histogram, Gaussian and Rayleigh distributions of TE operated noise PWP coefficients of noisy speech at an SNR of 15 dB

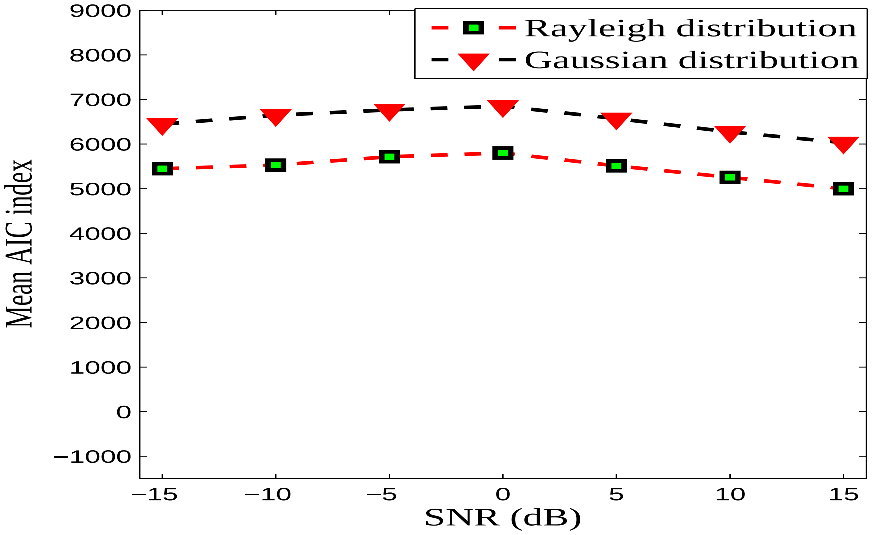


Fig. 5.7: Mean values of AIC index of TE operated PWP coefficients of noisy speech assuming Gaussian and Rayleigh distributions

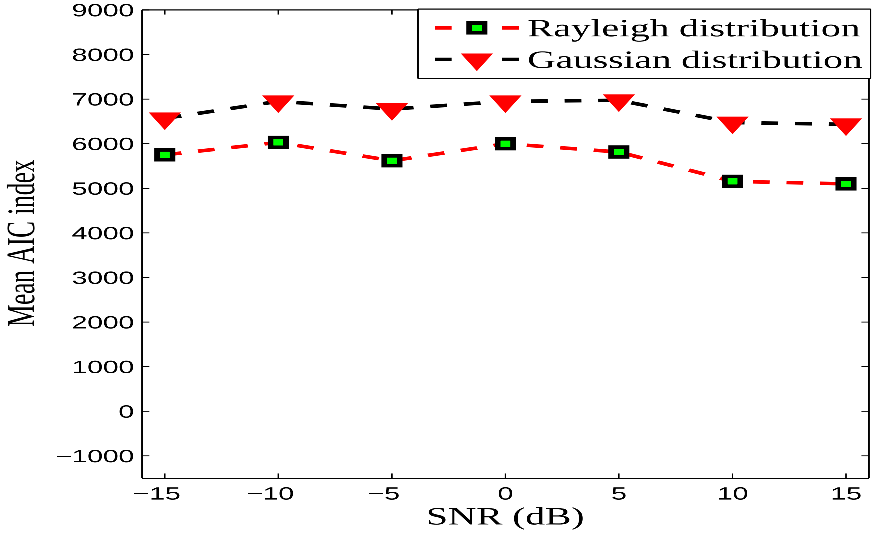


Fig. 5.8: Mean values of AIC index of TE operated PWP coefficients of noise assuming Gaussian and Rayleigh distributions

$$p_i(t_{k,m}) = \frac{x}{2\sigma_s^2} \times exp(-\frac{x}{\sigma_s^2}) \tag{5.1}$$

Where σ_s^2 represents the power of $t_{k,m}$ of noisy speech.

Letting σ_r^2 as the power of $t_{k,m}$ of clean speech and σ_n^2 as the power of $t_{k,m}$ of noise and using the fact $\sigma_s^2 = \sigma_r^2 + \sigma_n^2$, we can write

$$p_i(t_{k,m}) = \frac{x}{2(\sigma_r^2 + \sigma_n^2)} \times exp(-\frac{x}{\sigma_r^2 + \sigma_n^2}) \tag{5.2}$$

Following (5.1) in a similar way, Rayleigh pdf for $q_i(t_{k,m})$ can also be written as

$$q_i(t_{k,m}) = \frac{x}{2\sigma_n^2} \times exp(-\frac{x}{\sigma_n^2}) \tag{5.3}$$

By substituting (5.2) and (7.6) in (3.6), we obtain

$$\int_1^\lambda [\frac{x}{2\sigma_s^2} \times exp(-\frac{x}{\sigma_s^2}) - \frac{x}{2\sigma_n^2} \times exp(-\frac{x}{\sigma_n^2})]I_1 dx = 0 \tag{5.4}$$

where I_1 is defined as,

$$I_1 = ln(\frac{\sigma_s^2}{\sigma_n^2}) \times exp(-\frac{x}{\sigma_s^2} + \frac{x}{\sigma_n^2})$$

By solving (5.4), value of λ can be derived as

$$\lambda(k) = \sqrt{\frac{2}{\gamma_k}(1 + \frac{1}{\gamma_k})ln(1 + \gamma_k)} \tag{5.5}$$

where γ_k is segmental SNR of subband k defined as

$$\gamma_k = \frac{\sigma_r^2(k)}{\sigma_n^2(k)} \tag{5.6}$$

The proposed threshold $\lambda(k)$ in (5.5) derived assuming Rayleigh pdf is compared with that obtained assuming gaussian pdf given by

$$\lambda(k) = \frac{\sigma_n(k)}{\sqrt{\gamma_k}} \sqrt{2(\gamma_k + \gamma_k^2)} \times ln(\sqrt{1 + \frac{1}{\gamma_k}}) \tag{5.7}$$

in fig.5.9. This figure shows that the pattern of the threshold value is similar for both the pdfs at high as well as low SNRs. In terms of value, although Rayleigh pdf shows slightly lower values at high SNRs, but the threshold values are much lower than that of the gaussian pdf specially at low SNRs. Therefore, the threshold derived from the Rayleigh pdf offers less chance of removing speech coefficients while performing thresholding operation not only at high SNR but also at difficult low SNRs.

The proposed threshold as derived in (5.5) is high for higher noise power and low for lower noise power thus is adaptive to noise power of different subbands. In this method, voice activity detector is not needed as the threshold is automatically adapted to the silent and speech frames. At a silent frame, since noise power is significantly higher than the signal power, the proposed threshold results in a higher value as seen from (5.5). Such a value imposes more coefficients to be thresholded thus removing noise coefficients completely at subbands of a silent frame. Note that, in this paper, noise is estimated using Improved Minima Controlled Recursive Averaging (IMCRA) method [35].

On computing the threshold value as obtained above, the thresholding function as proposed in chapter 4 is employed on the PWP coefficients. The enhanced speech frame is synthesized by performing the inverse PWP transformation PWP^{-1} on the resulting thresholded PWP coefficients and The final enhanced speech signal is reconstructed by using the standard overlap-and-add method.

5.1 Results Considering Rayleigh Statistical Model

In this Section, a number of simulations is carried out with the same simulation conditions as described in chapter 3 to evaluate the performance of the proposed

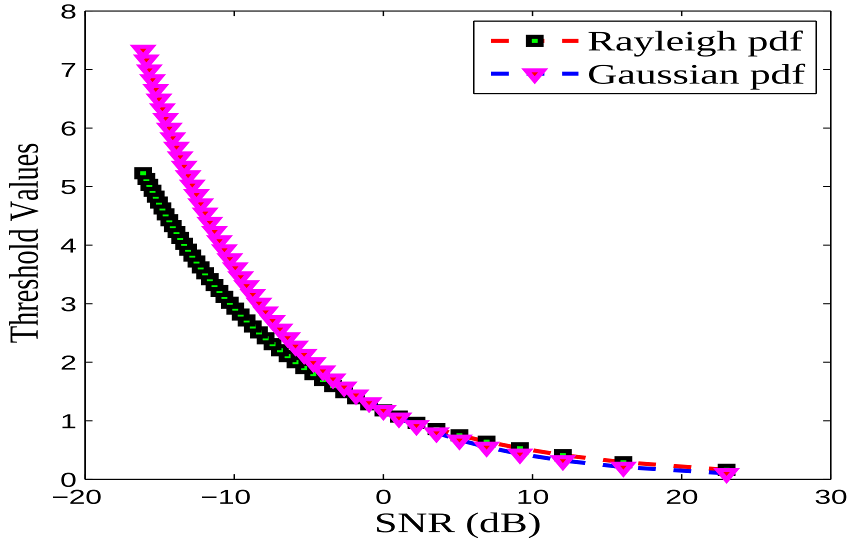


Fig. 5.9: Comparison of threshold values with respect to SNR for Rayleigh and Gaussian pdfs

method considering Rayleigh statistical model. Same comparison metrics are used to compare the proposed method with the previously mentioned comparison methods.

5.1.1 Objective Evaluation

Results for Speech signals with Car Noise

SNRSeg improvement, PESQ and WSS for speech signals corrupted with car noise for Universal, SMPO and proposed methods are shown in Fig.7.10, Table 5.1 and Fig.7.11.

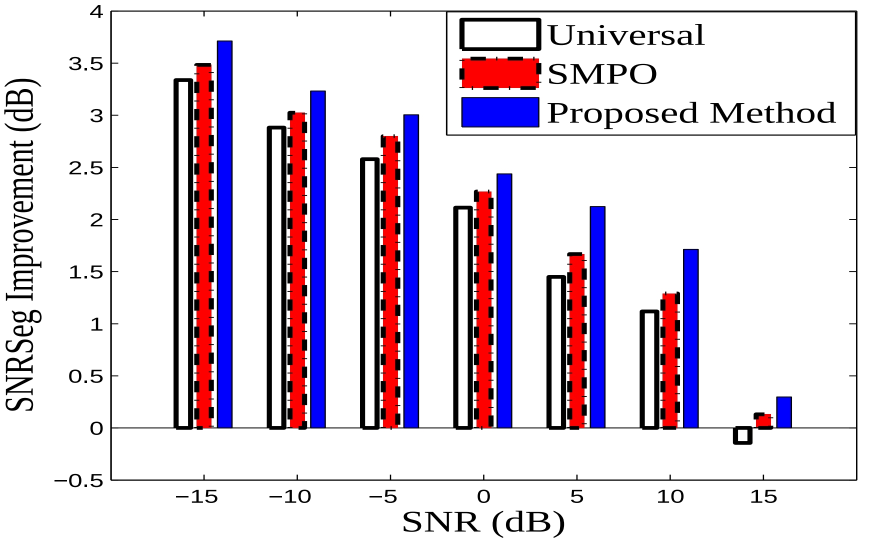


Fig. 5.10: SNRSeg Improvement for different methods in car noise

Table 5.1: PESQ for different methods in car noise

SNR(dB)	SMPO	Universal	Proposed Method
-15	1.15	1.16	1.21
-10	1.37	1.23	1.38
-5	1.51	1.32	1.54
0	1.69	1.43	1.71
5	2.07	1.69	2.14
10	2.38	1.93	2.48
15	2.60	2.14	2.83

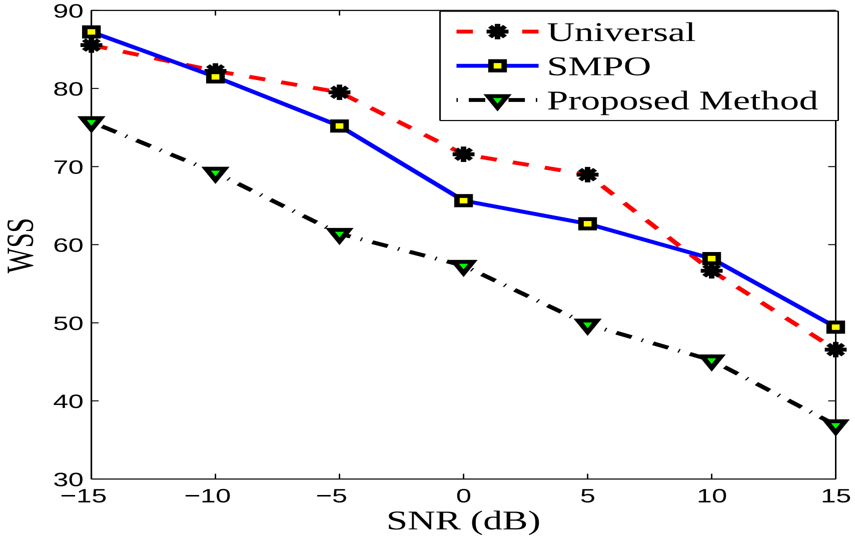


Fig. 5.11: WSS for different methods in car noise

In Figure 7.10, the performance of the proposed method is compared with that of the other methods at different levels of SNR for car noise in terms of Segmental SNR improvement. We see, the SNRSeg improvement increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields the highest SNRSeg improvement. Such larger values of SNRSeg improvement at a low level of SNR attest the capability of the proposed method in producing enhanced speech with better quality for speech severely corrupted by car noise.

In Table 5.1, it can be seen that at a low level of SNR, such as $-15dB$, all the methods show lower values of PESQ scores, whereas the PESQ score is much higher, as expected, for the proposed method. The proposed method also yields larger PESQ scores compared to that of the other methods at higher levels of SNR. Since, at a particular SNR, a higher PESQ score indicates a better speech quality, the proposed method is indeed better in performance in the presence of a car noise.

Fig.7.11 represents the WSS values as a function of SNR for the proposed method and that for the other methods. As shown in the figure, the WSS values resulting from all other methods are relatively larger for a wide range of SNR levels, whereas the proposed method is capable of producing enhanced speech with better quality as it gives lower values of WSS even at a low SNR of $-15dB$.

Results for Speech signals with Multi-talker Babble Noise

SNRSeg improvement, PESQ and WSS for speech signals corrupted with babble noise for Universal, SMPO and proposed methods are shown in Fig.7.12, 7.14 and 7.13, respectively.

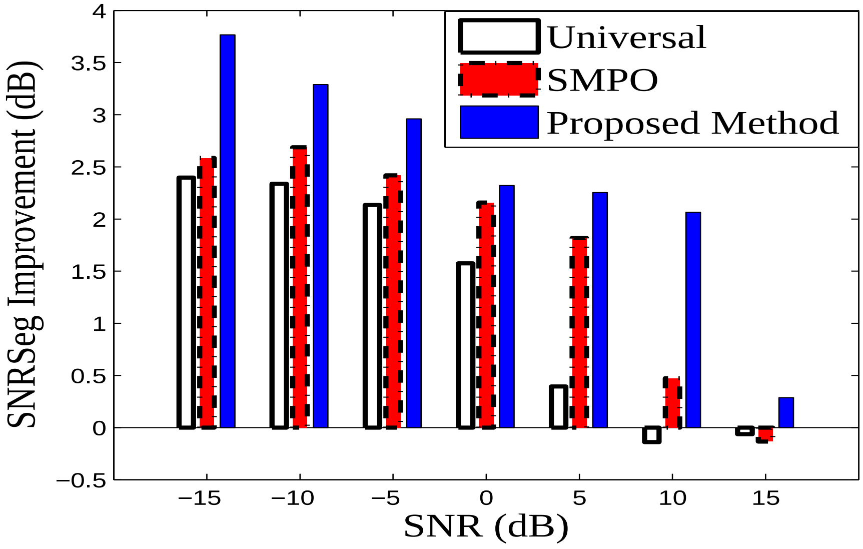


Fig. 5.12: SNRSeg Improvement for different methods in babble noise

In Fig. 7.12, it can be seen that at a low level of SNR of $-15dB$, the proposed method provides a SNRSeg improvement that is significantly higher than that of the methods of comparison. The proposed method still shows better performance in terms of SNRSeg improvement for higher SNRs also.

For speech corrupted with babble noise, in Fig.7.13, the mean values of PESQ with standard deviation obtained using the proposed method is plotted and compared with that of the other methods. From this plot, it is seen that over the whole SNR range considered, the proposed method continue to provide higher PESQ with almost non-overlapping standard deviation in the presence of babble noise.

The performance of the proposed method is compared with that of the other

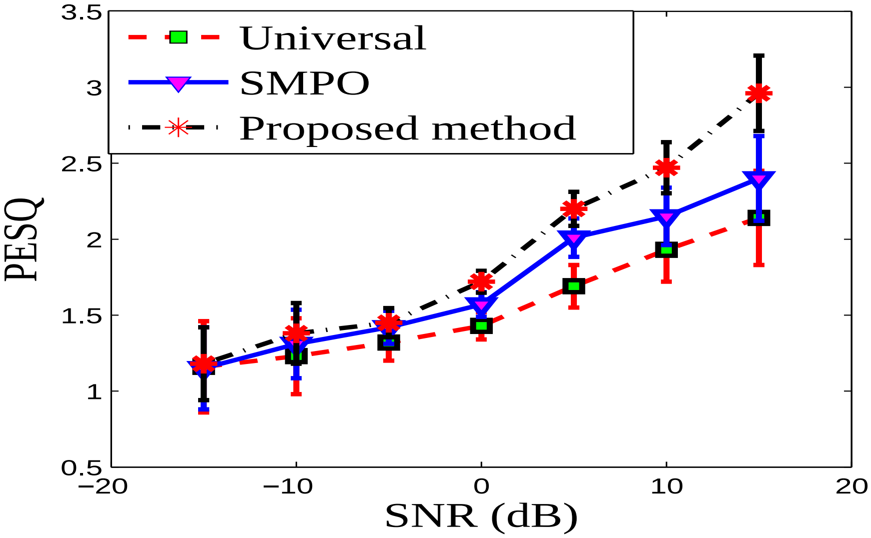


Fig. 5.13: PESQ for different methods in babble noise

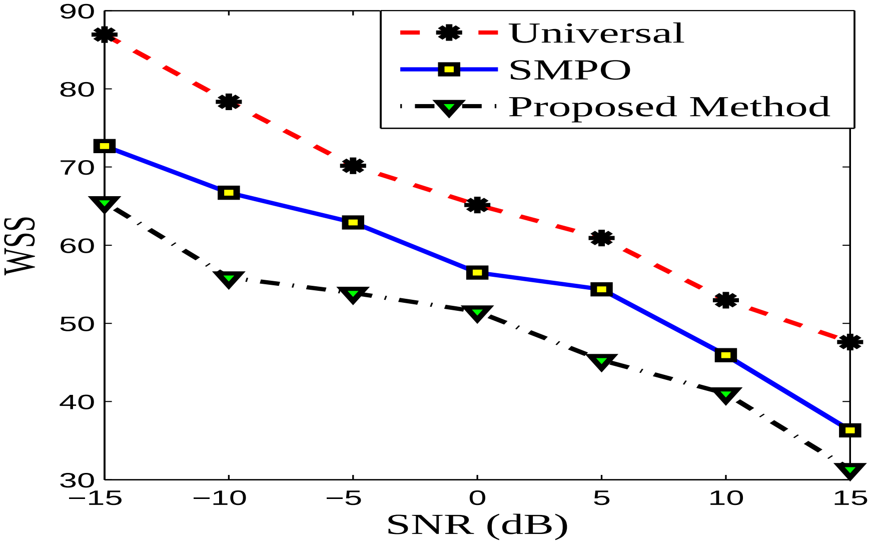


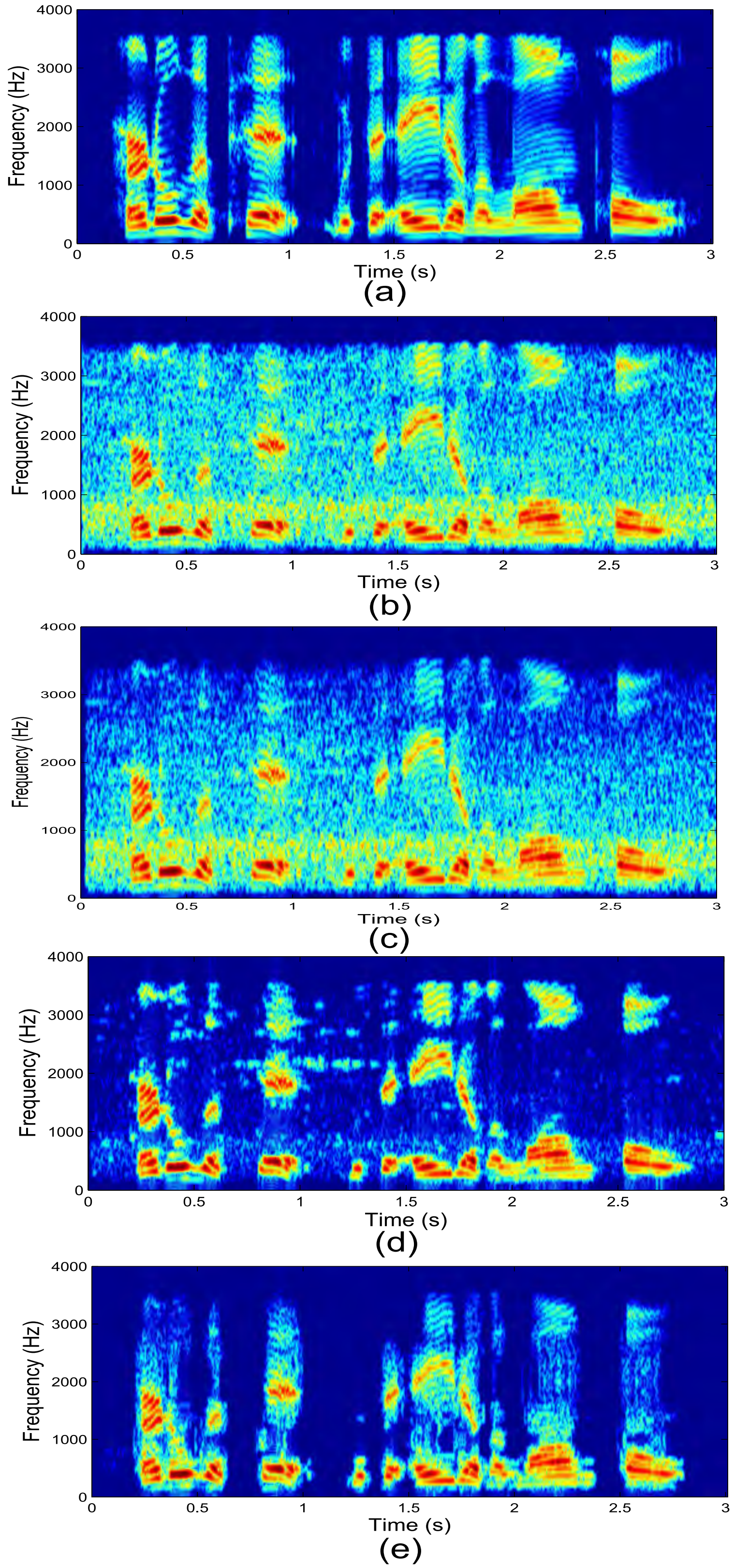
Fig. 5.14: WSS for different methods in babble noise

methods in terms of WSS in Fig.7.14 at different levels of SNRs in presence of babble noise. It is clearly seen from this figure that WSS increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields a WSS that is significantly lower than that of all other methods, which remains lower over the higher SNRs also.

5.1.2 Subjective Evaluation

In order to evaluate the subjective observation of the enhanced speech, spectrograms of the clean speech, the noisy speech, and the enhanced speech signals obtained by using the proposed method and all other methods are presented in Fig. 7.15 for car noise corrupted speech at an SNR of 10 dB. It is evident from this figure that the harmonics are well preserved and the amount of distortion is greatly reduced in the proposed method. Thus, the spectrogram observations with lower distortion also validate our claim of better speech quality as obtained in our objective evaluations in terms of higher SNR improvement in dB, higher PESQ score and lower WSS in comparison to the other methods. Another set of spectrograms for babble noise corrupted speech at an SNR of 10 dB is also presented in Fig.7.16. This figure attests that the proposed method has a better efficacy in preserving speech harmonics even in case of babble noise.

Formal listening tests are also conducted, where ten listeners are allowed and arranged to perceptually evaluate the enhanced speech signals. A full set (thirty sentences) of the NOIZEUS corpus was processed by Universal, SMPO and proposed method for subjective evaluation at different SNRs. Subjective tests were performed according to ITU-T recommendation P.835 [42]. In this tests, a listener is instructed to successively attend and rate the enhanced speech signal based on (a) the speech signal alone using a scale of SIG (1 = very unnatural, 5 = very natural), (b) the background noise alone using a scale of background conspicuous/ intrusiveness (BAK) (1 = very conspicuous, very intrusive; 5 = not noticeable), and (c) the overall effect using the scale of the mean opinion score (OVRL) (1 = bad, 5 = excellent). More details about the testing methodology can be found in [44]. The mean scores of SIG, BAK, and OVRL scales for the three speech enhancement methods evaluated in the presence of car noise at an SNR of 5 dB are shown in Tables 5.2, 5.3, and 5.4. For the three methods evaluated using babble noise-corrupted speech at an SNR of 10 dB, the mean scores of SIG, BAK, and OVRL scales are also summarized in Tables 5.5, 5.6, and 5.7. The mean scores in the presence of both car and babble noises demonstrate that the lower signal distortion (i.e., higher SIG scores) and the lower noise distortion (i.e., higher BAK scores) are obtained with the proposed method relative to that obtained by Universal and SMPO methods in most of the conditions. It is also shown that a consistently better performance in OVRL scale is offered by the proposed method not only in car but also in babble noisy conditions at both SNR levels considered in comparison to that provided by



(a)

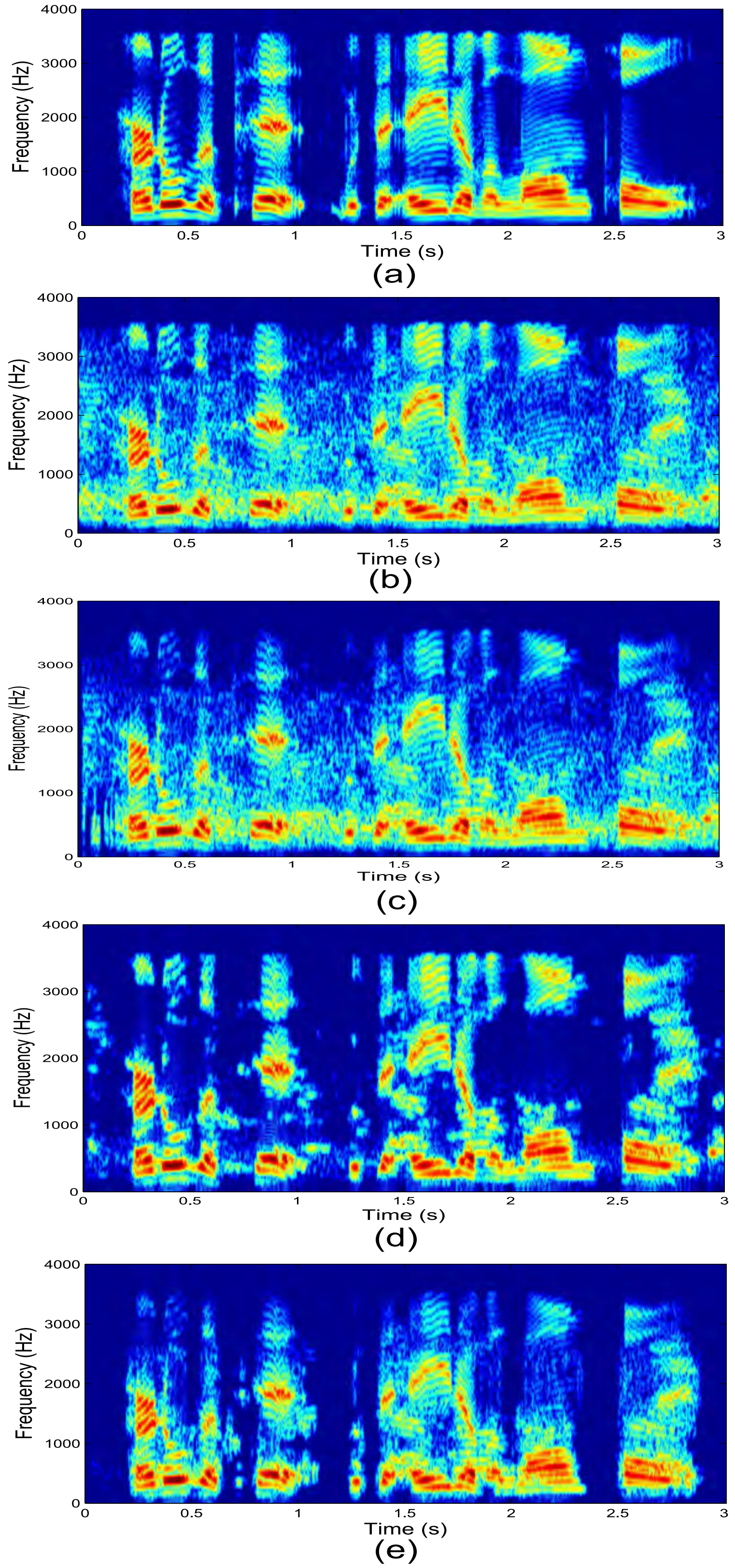
(b)

(c)

(d)

(e)

Fig. 5.15: Spectrograms of (a) Clean Signal (b) Noisy Signal with 10dB car noise; spectrograms of enhanced speech from (c) Universal method (d) SMPO method (e) Proposed Method



(a)

(b)

(c)

(d)

(e)

Fig. 5.16: Spectrograms of (a) Clean Signal (b) Noisy Signal with 10dB babble noise; spectrograms of enhanced speech from (c) Universal method (d) SMPO method (e) Proposed Method

Table 5.2: Mean Score of SIG scale for different methods in presence of car noise at 5 db

Listener	Universal	SMPO	Proposed Method
1	3.6	4.0	4.1
2	3.3	3.9	3.8
3	3.9	4.0	4.3
4	3.4	4.2	4.4
5	3.2	3.8	4.1
6	2.9	3.6	3.7
7	3.8	3.8	4.1
8	3.5	3.7	4.3
9	3.5	3.9	3.7
10	3.7	3.9	4.5

Table 5.3: Mean Score of BAK scale for different methods in presence of car noise at 5 db

Listener	Universal	SMPO	Proposed Method
1	4.0	4.5	4.9
2	4.3	4.9	4.8
3	4.2	4.4	4.5
4	4.4	4.7	4.7
5	4.2	4.8	4.6
6	3.9	4.6	4.8
7	3.8	3.9	4.5
8	4.4	4.6	4.4
9	3.5	3.8	4.6
10	4.2	4.5	4.9

all the methods mentioned above. Overall, it is found that the proposed method possesses the highest subjective sound quality in comparison to that of the other methods in case of different noises at various levels of SNRs.

5.2 Conclusions

In this paper, we developed a Rayleigh statistical model-based technique for the TE operated PWP coefficients of the noisy speech in order to obtain a suitable threshold value. By employing the proposed custom thresholding function designed based on the combination of μ-law and semisoft thresholding functions, the PWP coefficients of the noisy speech are thresholded in order to obtain an enhanced speech. It is shown through simulation results that the proposed method is able to yield consistently better results not only for car noise but also for multi-talker babble

Table 5.4: Mean Score of OVL scale for different methods in presence of car noise at 5 db

Listener	Universal	SMPO	Proposed Method
1	2.6	4.0	4.7
2	3.3	3.8	3.9
3	3.9	4.1	4.5
4	3.6	4.2	4.3
5	3.3	3.9	4.2
6	3.9	4.6	4.8
7	3.8	3.8	4.1
8	3.6	4.1	4.2
9	3.5	4.5	4.3
10	3.9	4.6	4.9

Table 5.5: Mean Score of SIG scale for different methods in presence of Babble noise at 5 db

Listener	Universal	SMPO	Proposed Method
1	3.6	4.0	4.4
2	3.3	3.9	3.8
3	3.9		4.0 4.3
4	3.4	4.2	4.8
5	3.2	3.8	4.4
6	2.9	3.6	4.0
7	3.8	3.8	4.3
8	3.4	3.6	4.2
9	3.5	3.9	3.8
10	3.7	3.8	3.8

Table 5.6: Mean Score of BAK scale for different methods in presence of Babble noise at 5 db

Listener	Universal	SMPO	Proposed Method
1	4.0	4.5	4.8
2	4.3	4.9	4.5
3	4.2		4.4 4.8
4	4.4	4.7	4.4
5	4.2	4.8	4.6
6	3.9	4.6	4.8
7	3.8	3.9	4.5
8	4.4	4.6	4.6
9	3.5	3.9	4.5
10	4.7	4.8	4.8

Table 5.7: Mean Score of OVL scale for different methods in presence of Babble noise at 5 db

Listener	Universal	SMPO	Proposed Method
1	2.6	4.0	4.2
2	3.3	3.8	3.9
3	3.9	4.1	4.4
4	3.6	4.2	4.5
5	3.3	3.9	4.2
6	3.9	4.6	4.8
7	3.8	3.8	4.3
8	3.6	4.1	4.4
9	3.5	4.5	4.8
10	3.9	4.8	4.7

noise corrupted speech signals in the sense of higher Segmental SNR Improvement in dB, higher output PESQ, and lower WSS values than those of the existing methods. The improvement in speech enhancement obtained by using the proposed method is also illustrated by the performance indicators, namely spectrogram outputs and scores in the formal subjective listening tests.

Chapter 6

Speech Enhancement Using Poisson Modeling of Teager Energy Operated Perceptual Wavelet Packet Coefficients

In this chapter, speech enhancement based on Poisson modeling of TE operated PWP coefficients is described [48]. An adaptive threshold is determined analytically using the Poisson model of TE operated PWP coefficients and then this threshold is imposed upon the PWP coefficients of noisy speech using custom thresholding function which is devised as a combination of μ-law and semisoft thresholding functions. Detail simulation has been performed to compare the proposed method with the state-of-the art speech enhancement techniques.

6.0.1 Proposed Poisson Distribution Model for TE Operated PWP Coefficients

Following discussion in chapter 3, As an alternative to formulate a pdf of the $t^j_{k,m}$ of speech, we can easily formulate the histogram of the $t^j_{k,m}$ and approximate the histogram by a reasonably close probability distribution function, namely Poisson distribution in place of Gaussian distribution [40]. For the $t_{k,m}$s in a subband of a noisy speech frame, the empirical histogram along with the Gaussian and the Poisson distributions are superimposed in Fig. 6.4, 6.5 and 6.6 in presence of car noise at SNRs of -15, 0 and 15 dB. From this figure, it is obvious that Poisson distribution fits the empirical histogram better than the Gaussian distribution. Similar analysis results are obtained for empirical histogram, Gaussian and Poisson distribution of TE operated noise PWP coefficients. Such statistical matching between

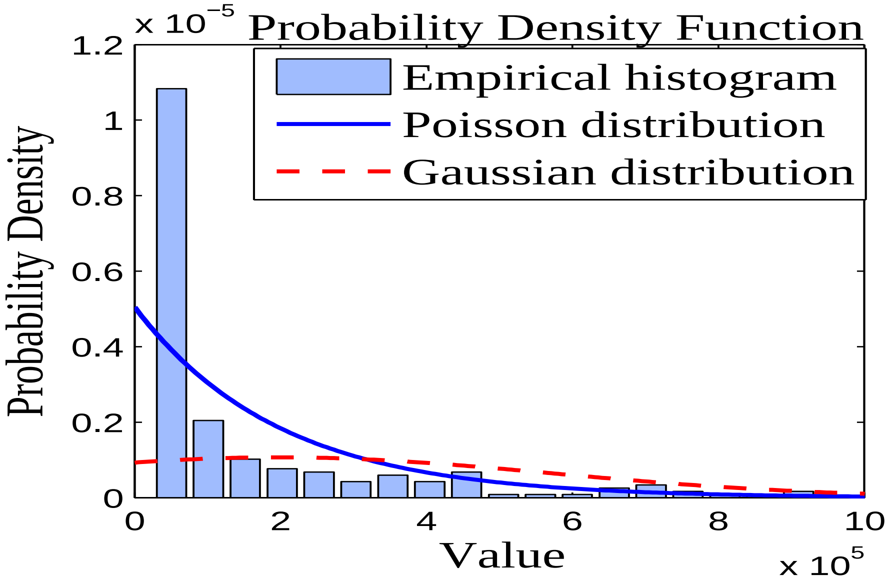


Fig. 6.1: Empirical histogram, Gaussian and Poisson distribution of TE operated PWP coefficients of noisy speech at SNR of -15 dB

the Gaussian and Poisson is also explained in terms of AIC index [46]. It can be noted from [46] that the more negative value of the AIC index indicates the more matching between two pdfs. Assuming Gaussian and Poisson distributions for $t_{k,m}$ in a subband of a noisy speech frame, mean values of AIC index obtained using different speech sentences are shown in fig. 6.7 for a range of SNR -15dB to 15dB in the presence of car noise. From fig. 6.7, it is clearly attested that the Poisson distribution offers better matching with the empirical histogram compared to the Gaussian distribution not only at SNR of 15dB but also at an SNR as low as -15dB. The plot representing the values of AIC index for the Gaussian and Poisson distributions of $t_{k,m}$ of noise at SNR level ranging from -15dB to 15dB is illustrated in fig. 6.8. This figure shows that AIC index for $t_{k,m}$ of noise continues to exhibit more negative values for Poisson distribution thus maintaining better pdf matching for a wide range of SNR. Therefore, we propose to approximate the histograms of $t_{k,m}$ of noisy speech, noise and clean speech by Poisson distribution and perform statistical modeling for calculating the threshold adaptive to different subbands.

6.0.2 Proposed Adaptive Threshold Calculation Assuming Poisson Distribution

Following the discussion in chapter 3, Poisson pdf for $p_i(t_{k,m})$ can be written as

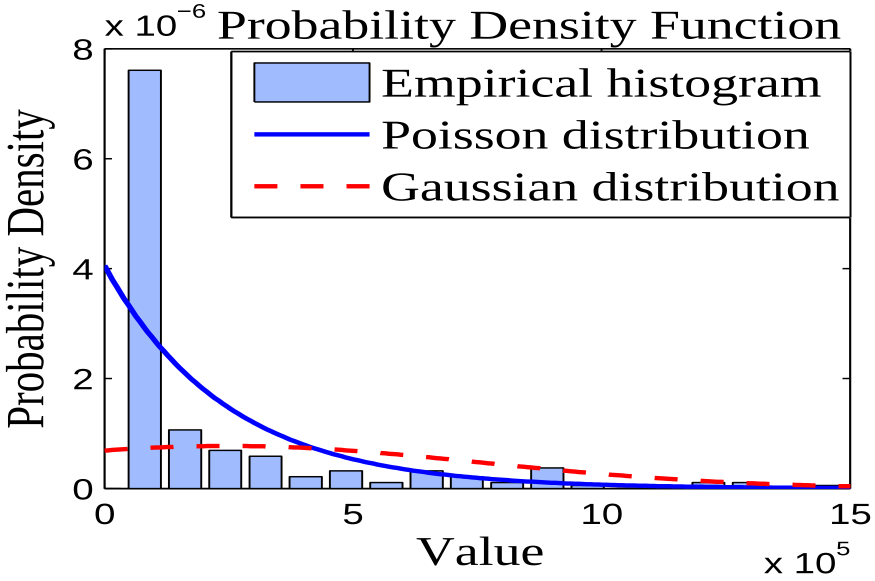


Fig. 6.2: Empirical histogram, Gaussian and Poisson distribution of TE operated PWP coefficients of noisy speech at SNR of 0 dB

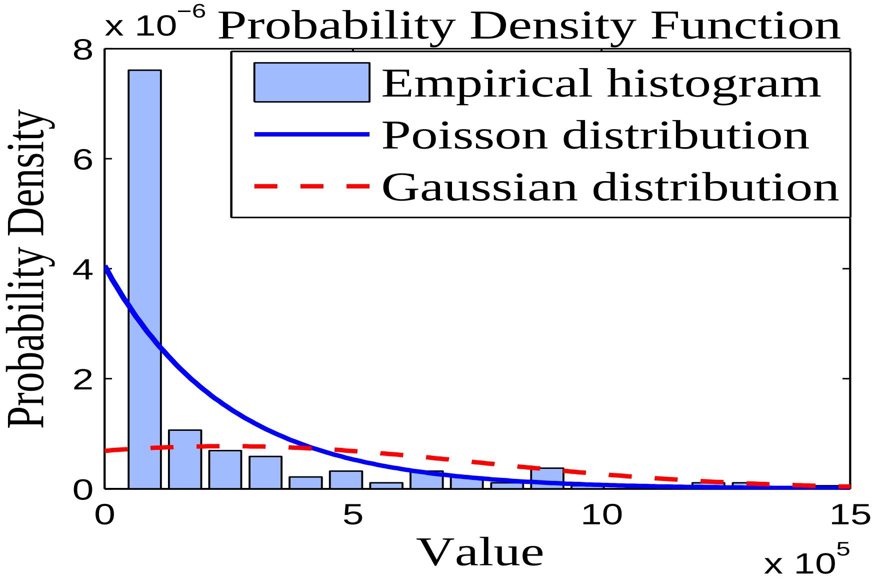


Fig. 6.3: Empirical histogram, Gaussian and Poisson distribution of TE operated PWP coefficients of noisy speech at SNR of 15 dB

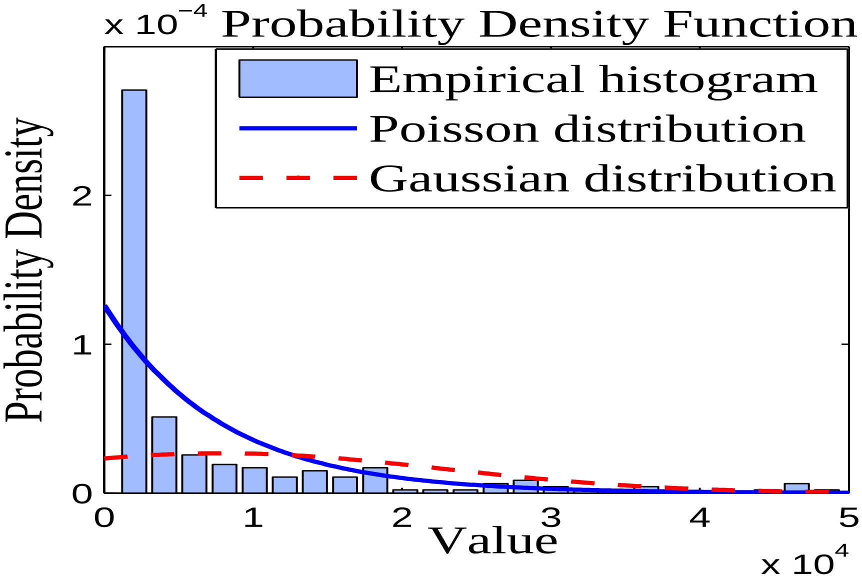


Fig. 6.4: Empirical histogram, Gaussian and Poisson distribution of TE operated PWP coefficients of noise at SNR of -15 dB

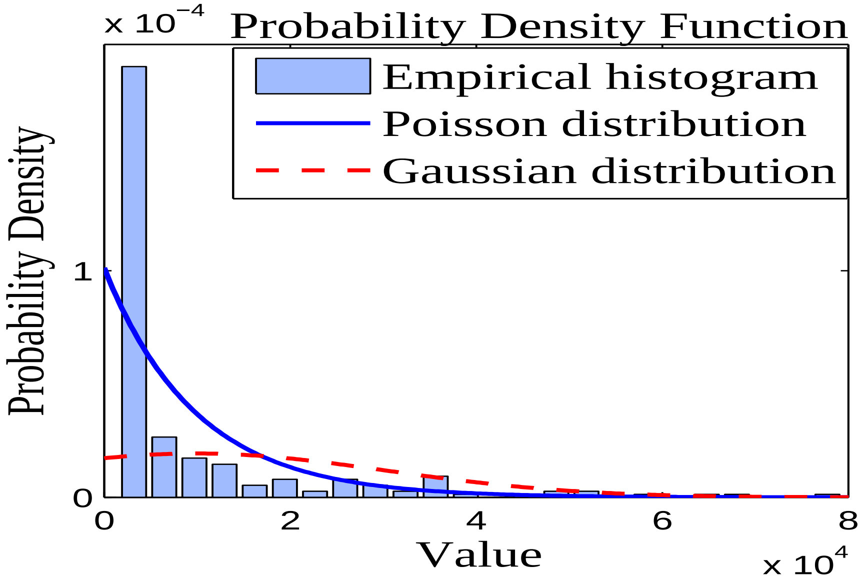


Fig. 6.5: Empirical histogram, Gaussian and Poisson distribution of TE operated PWP coefficients of noise at SNR of 0 dB

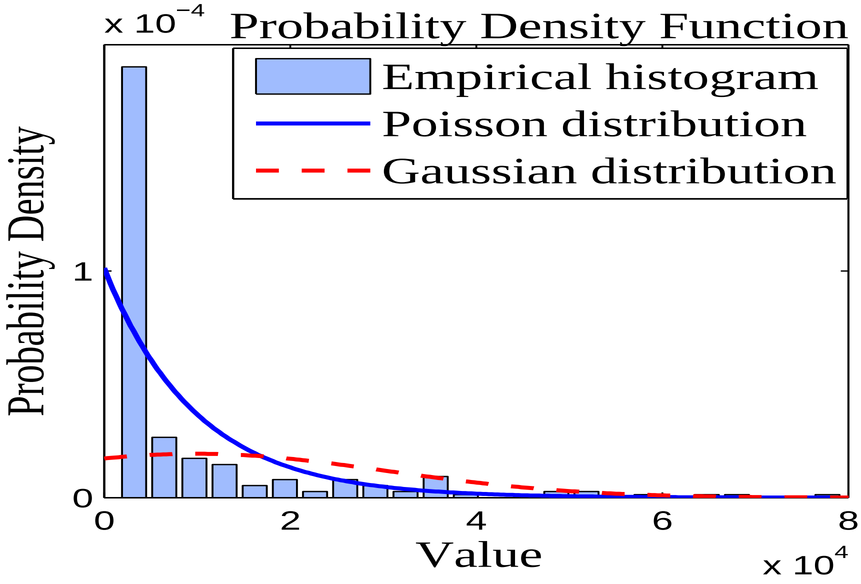


Fig. 6.6: Empirical histogram, Gaussian and Poisson distribution of TE operated PWP coefficients of noise at SNR of 15 dB

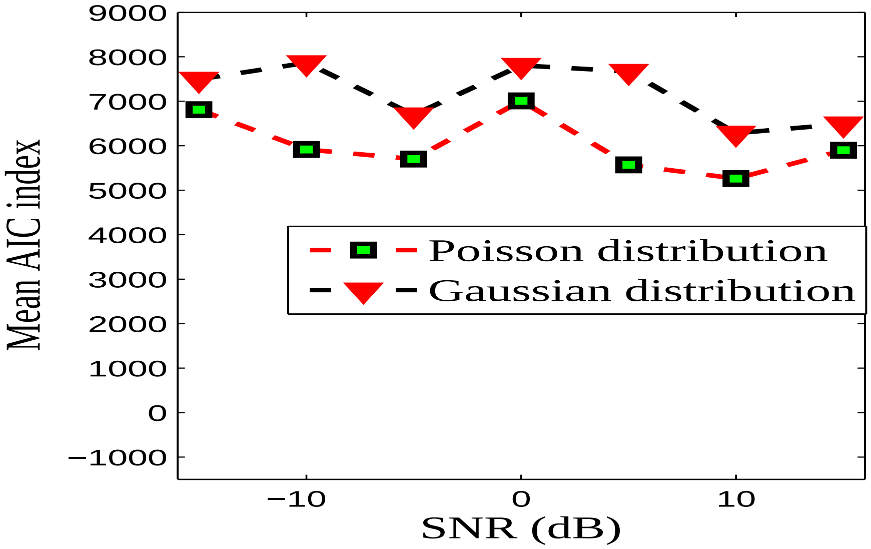


Fig. 6.7: Mean values of AIC index of TE operated PWP coeffiecients of noisy speech assuming Gaussian and Poisson distributions

$$p_i(t_{k,m}) = \frac{\sigma_s^{2x} e^{-\sigma_s^2}}{x!} \tag{6.1}$$

Where σ_s^2 represents the power of $t_{k,m}$ of noisy speech.

Letting σ_r^2 as the power of $t_{k,m}$ of clean speech and σ_n^2 as the power of $t_{k,m}$ of noise and using the fact $\sigma_s^2 = \sigma_r^2 + \sigma_n^2$, we can write

$$p_i(t_{k,m}) = \frac{\sqrt{\sigma_r^2 + \sigma_n^2}^{2x} e^{-(\sigma_r^2+\sigma_n^2)}}{x!} \tag{6.2}$$

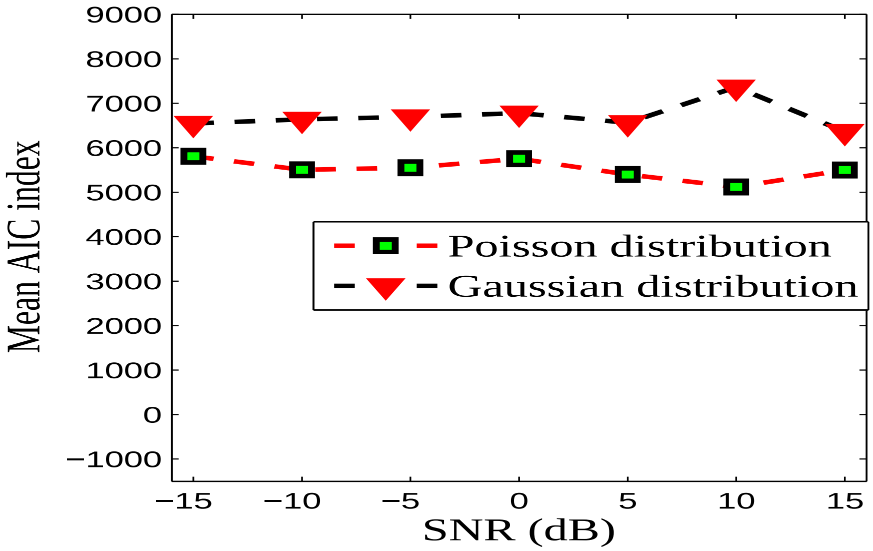


Fig. 6.8: Mean values of AIC index of TE operated PWP coeffiecients of noise assuming Gaussian and Poisson distributions

Following (7.1) in a similar way, laplace pdf for $q_i(t_{k,m})$ can also be written as

$$q_i(t_{k,m}) = \frac{\sigma_n^{2x} e^{-\sigma_n^2}}{x!} \tag{6.3}$$

By substituting (6.3) in (3.6), we obtain

$$\int_1^{\lambda} [\frac{\sigma_s^{2x} e^{-\sigma_s^2}}{x!} - \frac{\sigma_n^{2x} e^{-\sigma_n^2}}{x!}] I_1 dx = 0 \tag{6.4}$$

where I_1 is defined as,

$$I_1 = ln(\frac{\sigma_s^{2x}}{\sigma_n^{2x}}) \times e^{-\sigma_s^2 + \sigma_n^2}$$

By solving (6.4), value of λ can be derived as

$$\lambda(k) = \frac{\sqrt{\sigma_r^2(k)}}{ln(\frac{1}{1+\gamma_k})} \tag{6.5}$$

where γ_k is segmental SNR of subband k defined as

$$\gamma_k = \frac{\sigma_r^2(k)}{\sigma_n^2(k)} \tag{6.6}$$

The proposed threshold $\lambda(k)$ in (6.5) derived assuming Poisson pdf is compared with that obtained assuming Gaussian pdf given by

$$\lambda(k) = \frac{\sigma_n(k)}{\sqrt{\gamma_k}} \sqrt{2(\gamma_k + \gamma_k^2)} \times ln(\sqrt{1 + \frac{1}{\gamma_k}}) \tag{6.7}$$

in fig.6.9. This figure shows that the pattern of the threshold value is similar for both the pdfs at high as well as low SNRs. In terms of value, although Poisson pdf shows slightly lower values at high SNRs, but the threshold values are much lower than that of the Gaussian pdf specially at low SNRs. Therefore, the threshold derived from the Poisson pdf offers less chance of removing speech coefficients while performing thresholding operation not only at high SNR but also at difficult low SNRs.

The proposed threshold as derived in (6.5) is high for higher noise power and low for lower noise power thus is adaptive to noise power of different subbands. In this method, voice activity detector is not needed as the threshold is automatically adapted to the silent and speech frames. At a silent frame, since noise power is significantly higher than the signal power, the proposed threshold results in a higher value as seen from (6.5). Such a value imposes more coefficients to be thresholded thus removing noise coefficients completely at subbands of a silent frame. Note that, in this paper, noise is estimated using Improved Minima Controlled Recursive Averaging (IMCRA) method [35].

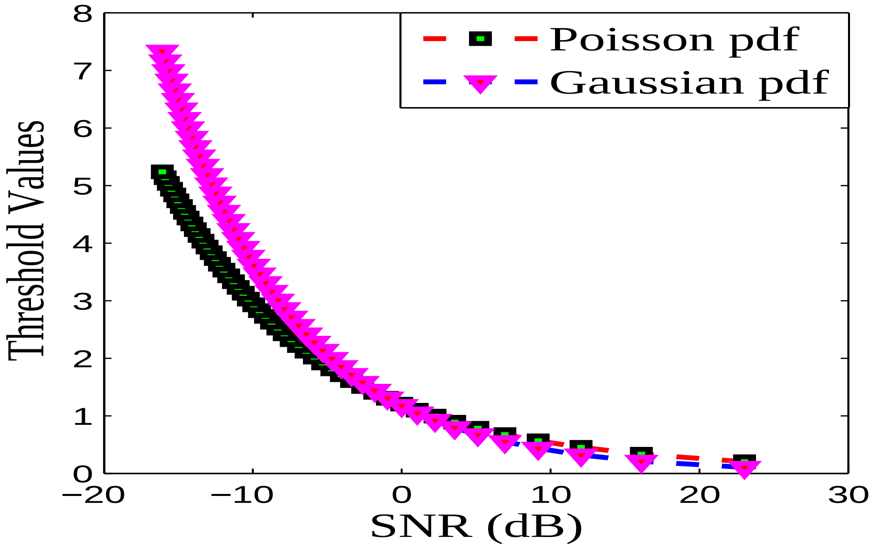


Fig. 6.9: Comparison of threshold values with respect to SNR for Poisson and Gaussian pdfs

On computing the threshold value as obtained above, the thresholding function as proposed in chapter 3 is employed on the PWP coefficients. Here the shape parameters are considered as constant values. The enhanced speech frame is synthesized by performing the inverse PWP transformation PWP^{-1} on the resulting thresholded PWP coefficients and The final enhanced speech signal is reconstructed by using the standard overlap-and-add method.

6.1 Results Considering Poisson Statistical Model

In this Section, a number of simulations is carried out with the same simulation conditions as described in chapter 3 to evaluate the performance of the proposed method considering Poisson statistical model. Same comparison metrics are used to compare the proposed method with the previously mentioned comparison methods.

6.1.1 Objective Evaluation

Results for Speech signals with Car Noise

SSNRSeg improvement, PESQ and WSS for speech signals corrupted with car noise for Universal, SMPO and proposed methods are shown in Fig.6.10, 6.12 and 6.11, respectively.

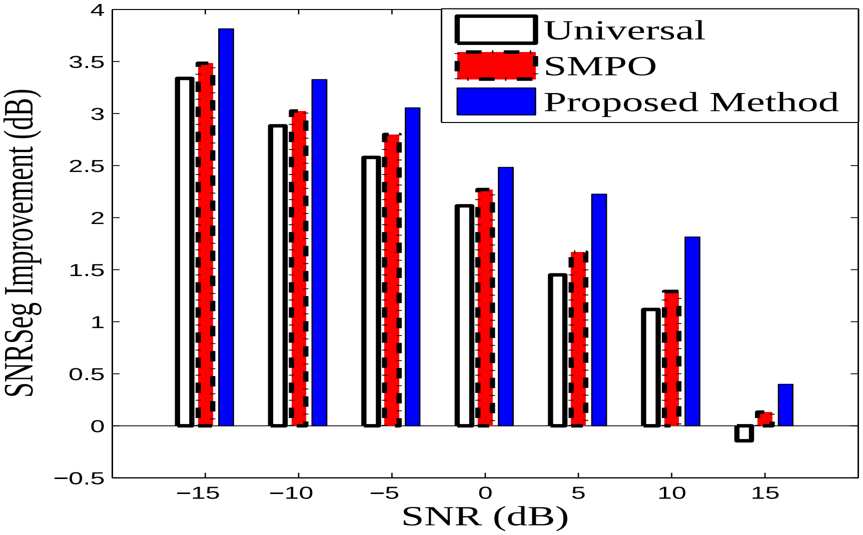


Fig. 6.10: SNRSeg Improvement for different methods in Car Noise assuming Poisson distribution for the proposed method

In Fig. 6.10, it can be seen that at a low level of SNR of $-15dB$, the proposed method provides a SNRSeg improvement that is significantly higher than that of the methods of comparison. The proposed method still shows better performance in terms of SNRSeg improvement for higher SNRs also.

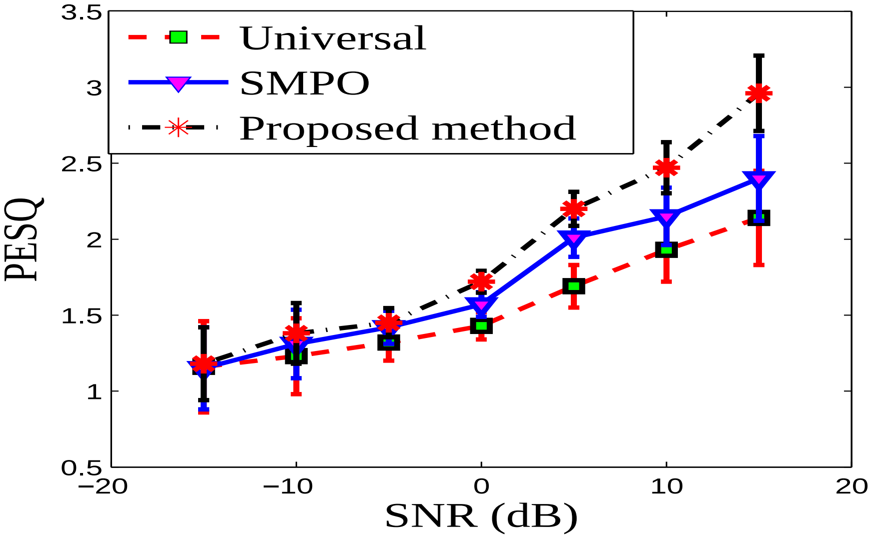


Fig. 6.11: PESQ for different methods in Car Noise assuming Poisson distribution for the proposed method

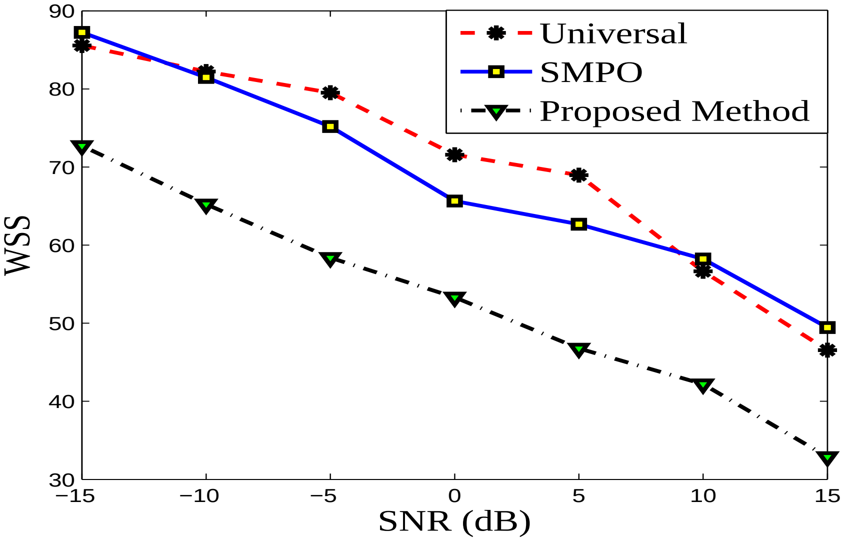


Fig. 6.12: WSS for different methods in Car Noise assuming Poisson distribution for the proposed method

For speech corrupted with babble noise, in Fig.6.11, the mean values of PESQ with standard deviation obtained using the proposed method is plotted and compared with that of the other methods. From this plot, it is seen that over the whole SNR range considered, the proposed method continue to provide higher PESQ with almost non-overlapping standard deviation in the presence of babble noise.

The performance of the proposed method is compared with that of the other methods in terms of WSS in Fig.6.12 at different levels of SNRs in presence of babble

Table 6.1: PESQ for different methods in Babble Noise assuming Poisson distribution for the proposed method

SNR(dB)	SMPO	Universal	Proposed Method (Poisson)
-15	1.26	1.13	1.27
-10	1.43	1.27	1.44
-5	1.61	1.41	1.61
0	1.77	1.53	1.82
5	2.12	1.77	2.14
10	2.43	1.97	2.48
15	2.66	2.15	2.81

noise. It is clearly seen from this figure that WSS increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields a WSS that is significantly lower than that of all other methods, which remains lower over the higher SNRs also.

Results for Speech signals with Multi-talker Babble Noise

SNRSeg, WSS and PESQ scores for speech signals corrupted with babble noise for Universal, SMPO and proposed method are shown in Fig.6.13, 6.14 and Table. 6.1.

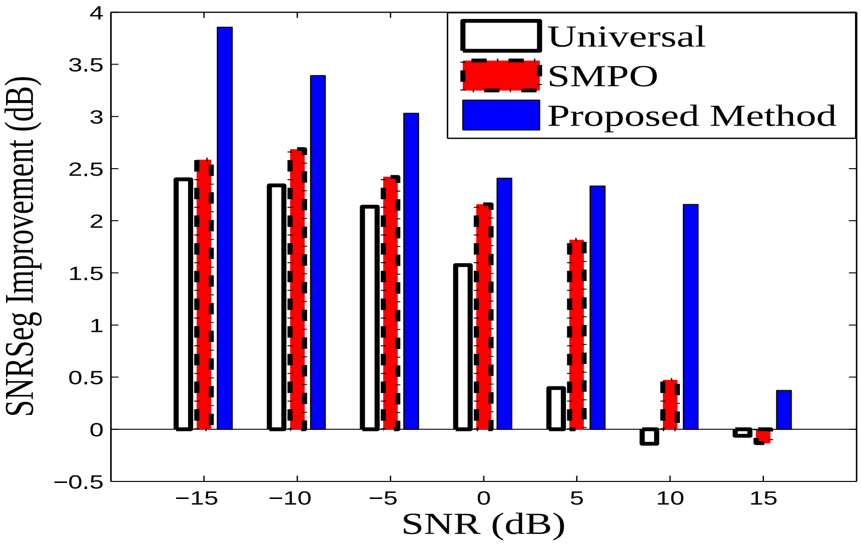


Fig. 6.13: SNRSeg Improvement for different methods in Babble Noise assuming Poisson distribution for the proposed method

In Figure 6.13, the performance of the proposed method is compared with that of the other methods at different levels of SNR for babble noise in terms of Segmental SNR inprovement. We see, the SNRSeg improvement increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields the highest SNRSeg improvement. Such larger values of SNRSeg improvement at a low level of SNR attest the capability

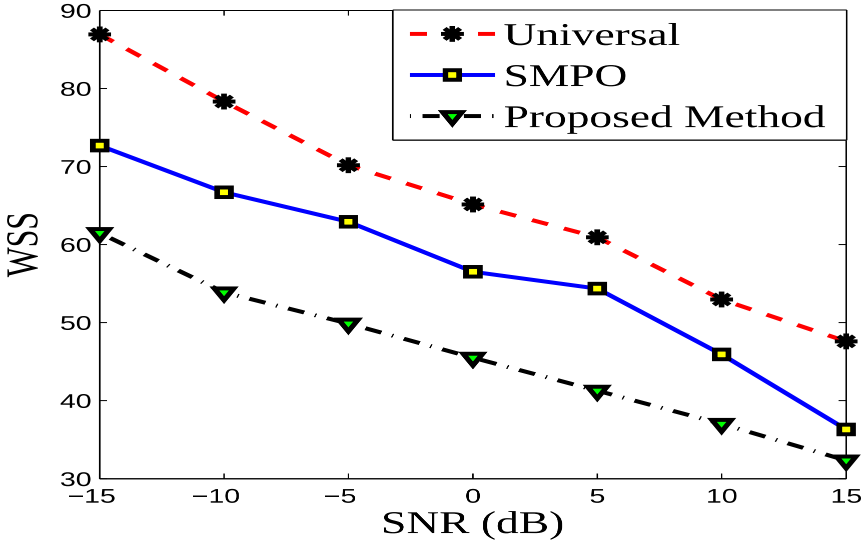


Fig. 6.14: WSS for different methods in Babble Noise assuming Poisson distribution for the proposed method

of the proposed method in producing enhanced speech with better quality for speech severely corrupted by car noise.

In Table.6.1, it can be seen that at a low level of SNR, such as $-15dB$, all the methods show lower values of PESQ scores, whereas the PESQ score is much higher, as expected, for the proposed method. The proposed method also yields larger PESQ scores compared to that of the other methods at higher levels of SNR. Since, at a particular SNR, a higher PESQ score indicates a better speech quality, the proposed method is indeed better in performance in the presence of a car noise.

Fig.6.14 represents the WSS values as a function of SNR for the proposed method and that for the other methods. As shown in the figure, the WSS values resulting from all other methods are relatively larger for a wide range of SNR levels, whereas the proposed method is capable of producing enhanced speech with better quality as it gives lower values of WSS even at a low SNR of $-15dB$.

6.1.2 Subjective Evaluation

In order to evaluate the subjective observation of the enhanced speech, spectrograms of the clean speech, the noisy speech, and the enhanced speech signals obtained by using the proposed method and all other methods are presented in Fig. 6.15 for car noise corrupted speech at an SNR of 10 dB. It is evident from this figure that the harmonics are well preserved and the amount of distortion is greatly reduced in the

Table 6.2: Mean Score of SIG scale for different methods in presence of car noise at 5 db assuming Poisson distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Poisson)**
1	4.0	3.6	4.0
2	3.9	3.3	3.7
3	4.0	3.9	4.2
4	4.2	3.4	4.5
5	3.8	3.2	4.0
6	3.6	2.9	3.9
7	3.8	3.8	4.2
8	3.7	3.5	4.2
9	3.9	3.5	3.8
10	3.9	3.7	4.0

proposed method. Thus, the spectrogram observations with lower distortion also validate our claim of better speech quality as obtained in our objective evaluations in terms of higher SNR improvement in dB, higher PESQ score and lower WSS in comparison to the other methods. Another set of spectrograms for babble noise corrupted speech at an SNR of 10 dB is also presented in Fig.6.16. This figure attests that the proposed method has a better efficacy in preserving speech harmonics even in case of babble noise.

The mean scores of SIG, BAK, and OVRL scales for the three speech enhancement methods evaluated in the presence of car noise at an SNR of 5 dB are shown in Tables 6.2, 6.3, and 6.4. For the three methods examined using babble noise-corrupted speech at an SNR of 10 dB, the mean scores of SIG, BAK, and OVRL scales are summarized in Tables 6.5, 6.6, and 6.7. The mean scores in the presence of both car and babble noises demonstrate that the lower signal distortion (i.e., higher SIG scores) and the lower noise distortion (i.e., higher BAK scores) are obtained with the proposed method relative to that obtained by Universal and SMPO methods in most of the conditions. It is also shown that a consistently better performance in OVRL scale is offered by the proposed method not only in car but also in all other noisy conditions at both SNR levels of considered in comparison to that provided by all the methods mentioned above. Overall, it is found that the proposed method possesses the highest subjective sound quality in comparison to that of the other methods in case of different noises at various levels of SNR.

Table 6.3: Mean Score of BAK scale for different methods in presence of car noise at 5 db assuming Poisson distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Poisson)**
1	4.5	4.0	5.0
2	4.9	4.3	4.7
3	4.4	4.2	4.9
4	4.7	4.4	4.8
5	4.8	4.2	4.7
6	4.6	3.9	4.9
7	3.9	3.8	4.4
8	4.6	4.4	4.6
9	3.8	3.5	4.5
10	4.5	4.2	4.8

Table 6.4: Mean Score of OVL scale for different methods in presence of car noise at 5 db assuming Poisson distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Poisson)**
1	4.0	2.6	4.1
2	3.8	3.3	3.7
3	4.1	3.9	4.3
4	4.2	3.6	4.2
5	3.9	3.3	4.1
6	4.6	3.9	4.9
7	3.8	3.8	4.3
8	4.1	3.6	4.2
9	4.5	3.5	4.7
10	4.6	3.9	4.8

Table 6.5: Mean Score of SIG scale for different methods in presence of Babble noise at 5 db assuming Poisson distribution for the proposed method

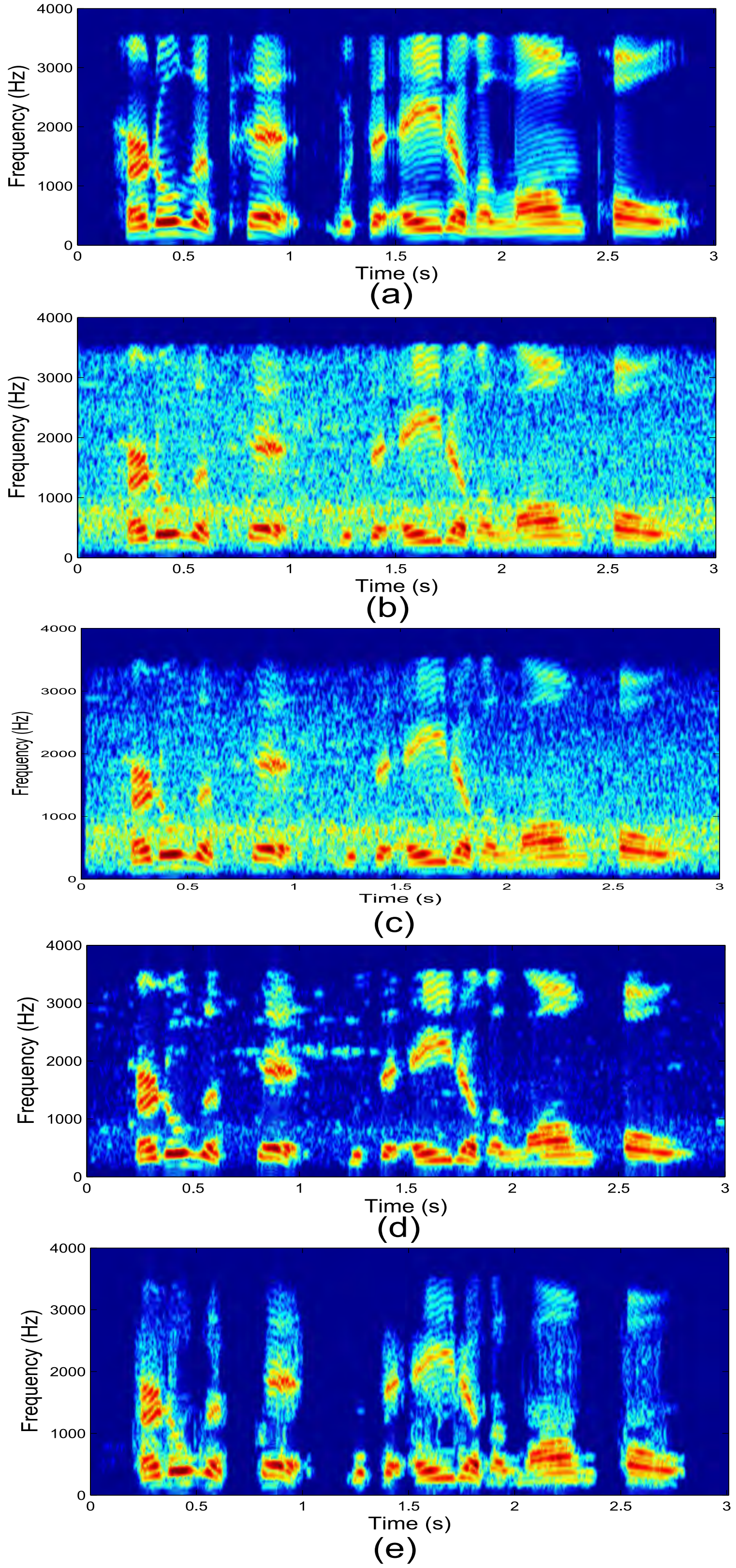
Listener	**SMPO**	**Universal**	**Proposed Method (Poisson)**
1	4.0	3.6	4.0
2	3.9	3.3	3.7
3	4.0	3.9	4.2
4	4.2	3.4	4.5
5	3.8	3.2	4.0
6	3.6	2.9	3.9
7	3.8	3.8	4.2
8	3.6	3.4	4.1
9	3.9	3.5	3.7
10	3.8	3.7	3.9

Table 6.6: Mean Score of BAK scale for different methods in presence of Babble noise at 5 db assuming Poisson distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Poisson)**
1	4.5	4.0	5.0
2	4.9	4.3	4.7
3	4.4	4.2	4.9
4	4.7	4.4	4.8
5	4.8	4.2	4.7
6	4.6	3.9	4.9
7	3.9	3.8	4.4
8	4.6	4.4	4.7
9	3.9	3.5	4.7
10	4.8	4.7	4.9

Table 6.7: Mean Score of OVL scale for different methods in presence of Babble noise at 5 db assuming Poisson distribution for the proposed method

Listener	**SMPO**	**Universal**	**Proposed Method (Poisson)**
1	4.0	2.6	4.1
2	3.8	3.3	3.7
3	4.1	3.9	4.3
4	4.2	3.6	4.2
5	3.9	3.3	4.1
6	4.6	3.9	4.9
7	3.8	3.8	4.3
8	4.1	3.6	4.2
9	4.5	3.5	4.7
10	4.8	3.9	4.9



(a)

(b)

(c)

(d)

(e)

Fig. 6.15: Spectogram of Output for noisy signal mixed with 10dB car noise for different methods (a) Clean Signal (b) Noisy Signal (c) SMPO (d) Universal (e) Proposed Method assuming Poisson distribution

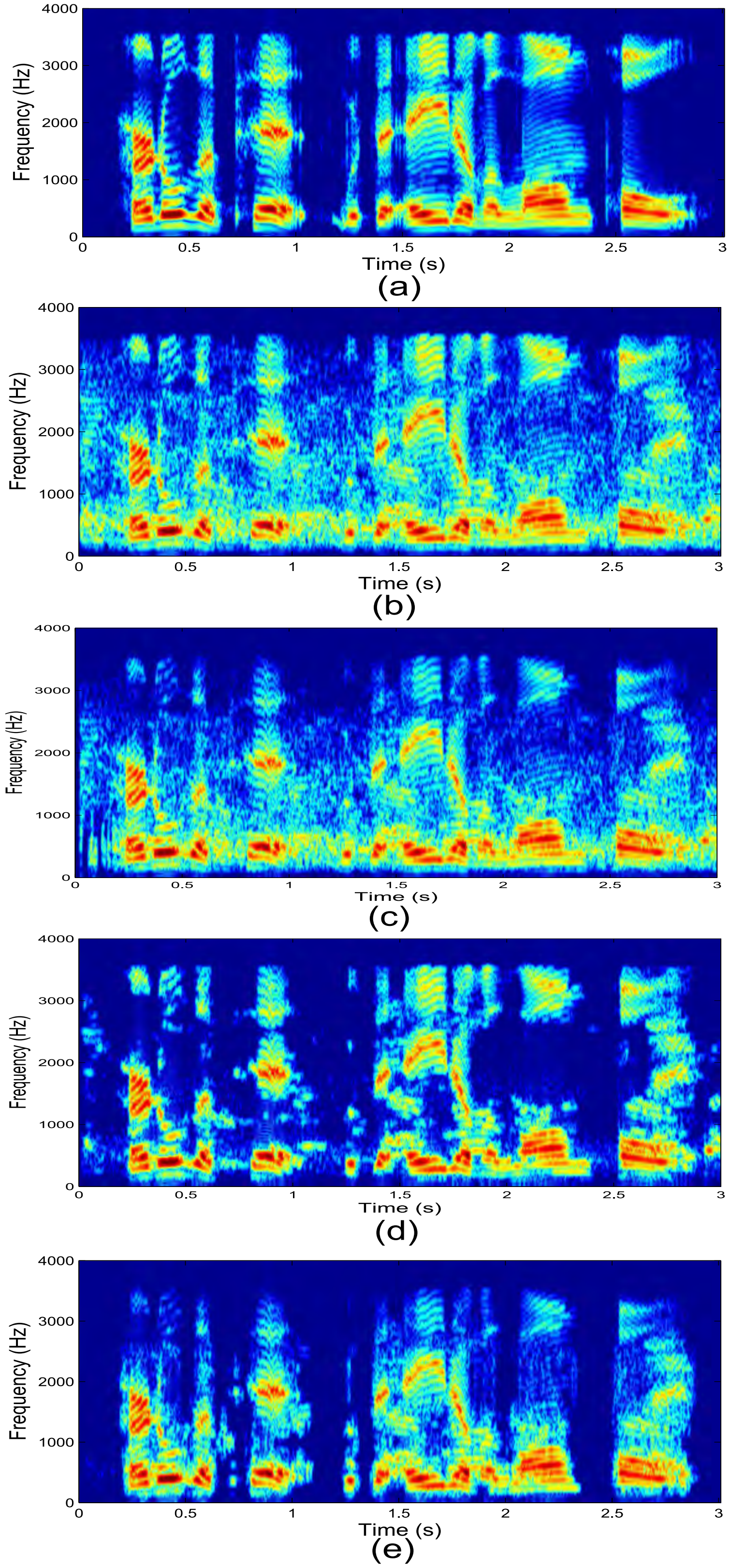


Fig. 6.16: Spectogram of Output for noisy signal mixed with 10dB babble noise for different methods (a) Clean Signal (b) Noisy Signal (c) SMPO (d) Universal (e) Proposed Method assuming Poisson distribution distribution

6.2 Conclusion

To solve the problems of speech enhancement, an improved perceptual wavelet packet based approach using the Poisson pdf of Teager Energy Operated wavelet Packet coefficients has been presented in this paper. We incorporated a statistical model-based technique with teager energy operator of the wavelet packet coefficients to obtain a suitable threshold using symmetric K-L divergence. For solving the equation of pdf's, we choose Poisson distribution as an acceptable pdf for noisy speech, clean speech and noise TE operated PWP coefficients in each sub-band. Unlike the unique threshold based method, the threshold value here is adapted based on the speech and silence segments. Then, by employing the proposed custom thresholding function, the PWP coefficients of the noisy speech are thresholded in order to obtain a cleaner speech. Simulation results show that the proposed method yields consistently better results in the sense of higher output SNR in dB, higher output PESQ, and lower WSS values than those of the existing methods.

Chapter 7

Speech Enhancement Using Student t Modeling of Teager Energy Operated Perceptual Wavelet Packet Coefficients

In this chapter, speech enhancement based on Student t modeling of TE operated PWP coefficients is described [49]. An adaptive threshold is determined analytically using the Student t model of TE operated PWP coefficients and then this threshold is imposed upon the PWP coefficients of noisy speech using pdf dependent custom thresholding function which is devised as a combination of modified hard and semisoft thresholding functions. Detail simulation has been performed to compare the proposed method with the state-of-the art speech enhancement techniques.

7.0.1 Proposed Student t Distribution Model for TE Operated PWP Coefficients

Following discussion in chapter 3, As an alternative to formulate a pdf of the of speech, we can easily formulate the histogram of its $t_{k,m}$ and can approximate the histogram by a reasonably close pdf namely gaussian and student t distribution. For the $t_{k,m}$s in a subband of a noisy speech frame, the empirical histogram along with the gaussian and the student t distributions are superimposed in Fig. 7.1,7.2 and 7.3 in presence of car noise at SNRs of -15, 0 and 15 dB. From this figure, it is obvious that Student t distribution fits the empirical histogram better than the Gaussian distribution. Similar analysis results are obtained for empirical histogram, gaussian and student t distribution of TE operated noise PWP coefficients at the same SNRs as used in Fig. 7.1,7.2 and 7.3 and are shown in Fig.7.4,7.5 and 7.6.

Such statistical matching between the gaussian and student t is also explained in terms of AIC index [46]. It can be noted from [46] that the more negative value of the AIC index indicates the more matching between two pdfs. Assuming gaussian and student t distributions for $t_{k,m}$ in a subband of a noisy speech frame, mean values of AIC index obtained using different speech sentences are shown in fig. 7.7 for a range of SNR -15dB to 15dB in the presence of car noise. From fig. 7.7, it is clearly attested that the student t distribution offers better matching with the empirical histogram compared to the gaussian distribution not only at SNR of 15dB but also at an SNR as low as -15dB. The plot representing the values of AIC index for the gaussian and student t distributions of $t_{k,m}$ of noise at SNR level ranging from -15dB to 15dB is illustrated in fig. 7.8. This figure shows that AIC index for $t_{k,m}$ of noise continues to exhibit more negative values for student t distribution thus maintaining better pdf matching for a wide range of SNR. Therefore, we propose to approximate the histograms of $t_{k,m}$ of noisy speech, noise and clean speech by student t distribution and perform statistical modeling for calculating the threshold adaptive to different subbands.

It is well known that the more negative the AIC index becomes, the more pdf matching it indicates. In Fig. 7.7, AIC indices assuming Gaussian as well as Student t distribution for TE operated PWPT coefficients for a subband of a noisy frame are shown for SNR of -15 dB to 15 dB. This figure also attests that Student t distribution offers better matching with the empirical data for all speech files not only at high SNR but also at low SNR as -15 dB. The plot of index AIC illustrated in Fig. 7.8 for Gaussian and Student t distribution of TE operated PWP coefficients at SNR level of -15dB to 15 dB continue to show more negative AIC values for Student t distribution maintaining better pdf matching for a wide level of SNR. Therefore, we are motivated in this research to perform statistical modeling of PWP coefficients via Student t distribution.

7.0.2 Proposed Adaptive Threshold Calculation assuming Student *t* distribution

Following the discussion in chapter 3, student t distribution pdf for $p_i(t_{k,m})$ can be written as

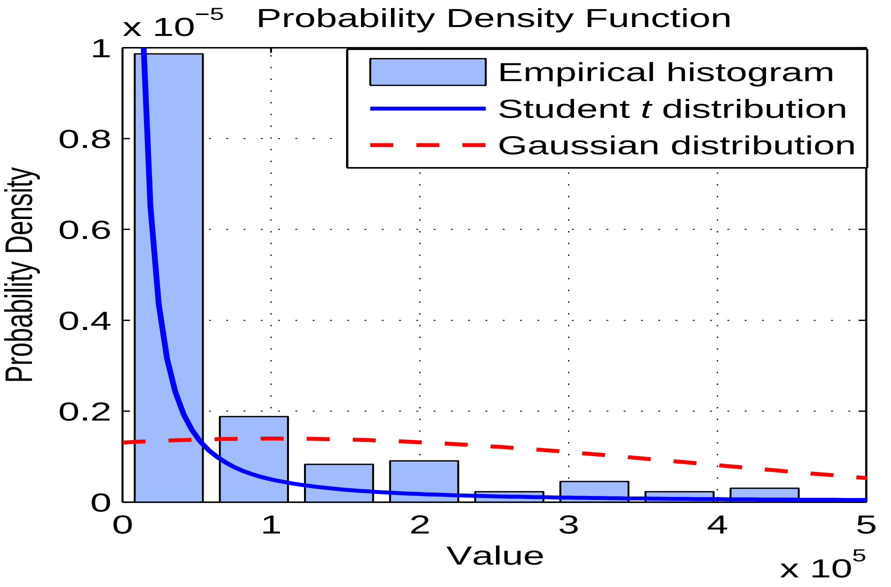


Fig. 7.1: Empirical histogram, Gaussian and Student t distribution of TE operated PWP coefficients of noisy speech at SNR of -15 dB

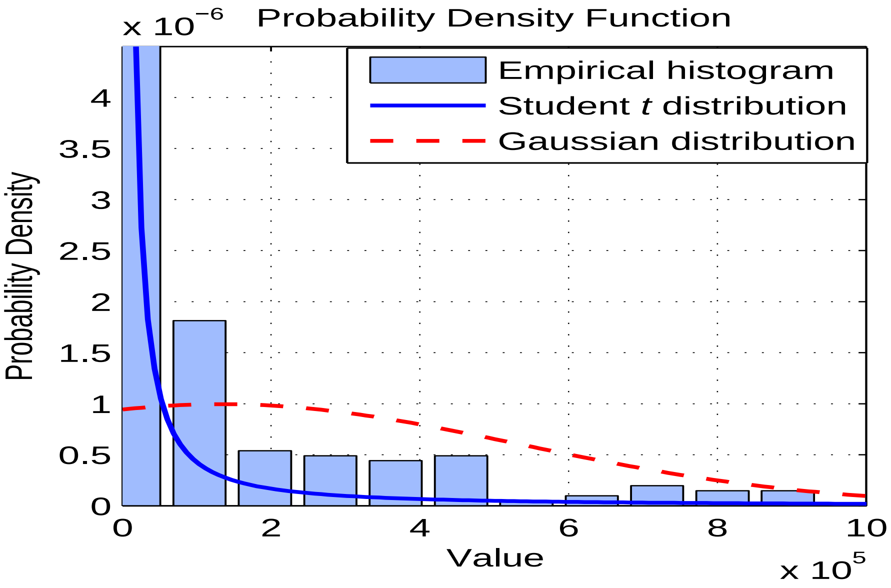


Fig. 7.2: Empirical histogram, Gaussian and Student t distribution of TE operated PWP coefficients of noisy speech at SNR of 0 dB

$$p_i(t_{k,m}) = \frac{\chi_1}{\sigma_s}(1 + \frac{1}{\nu - 2}\frac{x^2}{\sigma_s^2})^{-\frac{\nu+1}{2}}. \tag{7.1}$$

Where ν denotes the degree of freedom, σ_s^2 represents the power of $t_{k,m}$ of noisy speech and χ_1 is defined as

$$\chi_1 = \frac{\gamma(\frac{\nu+1}{2})}{\gamma(\frac{\nu}{2})\sqrt{\pi(\nu-2)}} \tag{7.2}$$

Letting σ_r^2 as the power of $t_{k,m}$ of clean speech and σ_n^2 as the power of $t_{k,m}$ of

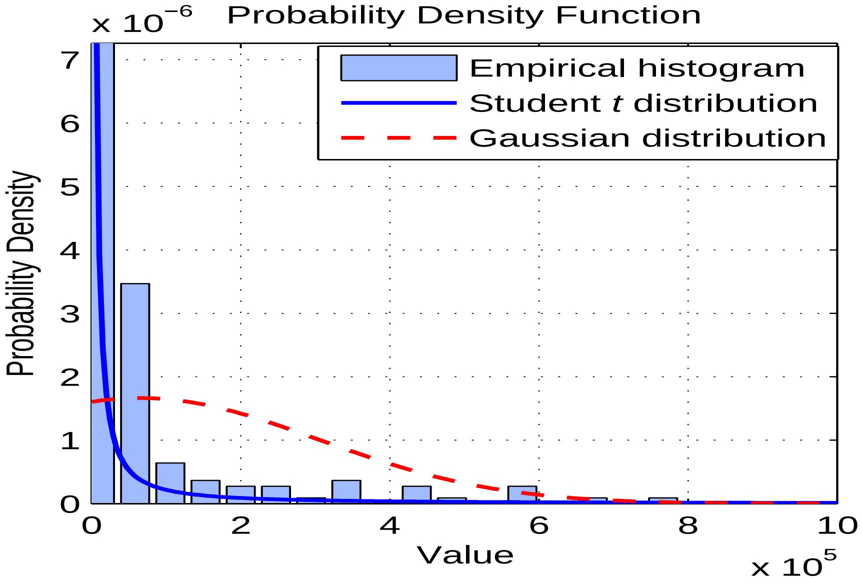


Fig. 7.3: Empirical histogram, Gaussian and Student t distribution of TE operated PWP coefficients of noisy speech at SNR of 15 dB

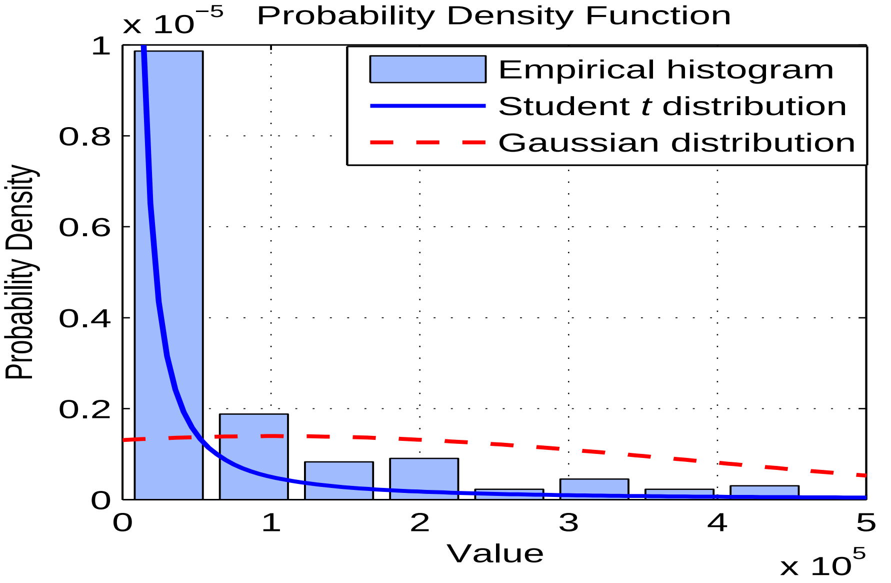


Fig. 7.4: Empirical histogram, Gaussian and Student t distribution of TE operated noise PWP coefficients at SNR of −15 dB

noise and using the fact $\sigma_s^2 = \sigma_r^2 + \sigma_n^2$, we can write

$$p_i(t_{k,m}) = \frac{\chi_1}{\sqrt{\sigma_r^2 + \sigma_n^2}}(1 + \frac{1}{\nu - 2}\frac{x^2}{\sigma_r^2 + \sigma_n^2})^{-\frac{\nu+1}{2}} \tag{7.3}$$

For $\frac{x^2}{\nu(\sigma_r^2+\sigma_n^2)} \ll 1$, Using binomial theorem, (7.3) can be approximated as

$$p_i(t_{k,m}) = \frac{\chi_1}{\sqrt{\sigma_r^2 + \sigma_n^2}}(1 - \chi_2\frac{x^2}{\sigma_r^2 + \sigma_n^2}) \tag{7.4}$$

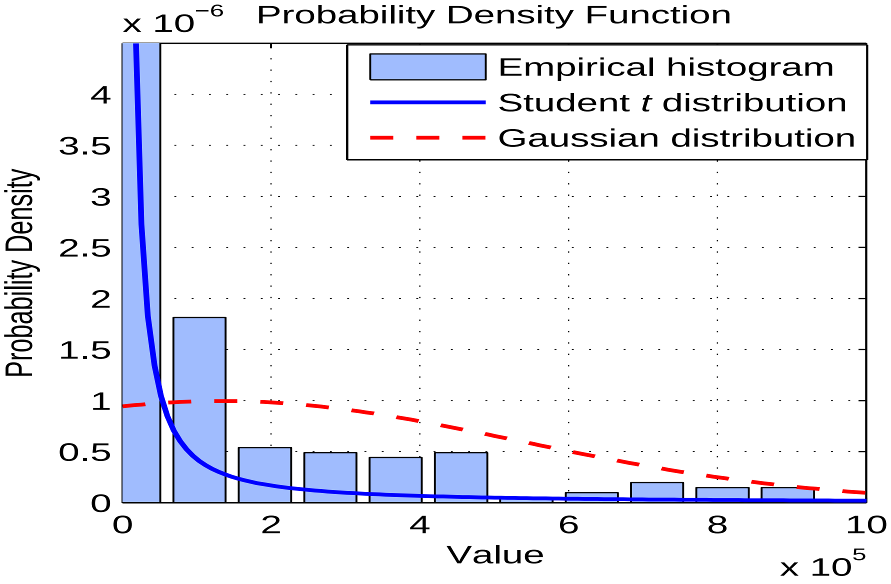


Fig. 7.5: Empirical histogram, Gaussian and Student t distribution of TE operated noise PWP coefficients at SNR of 0 dB

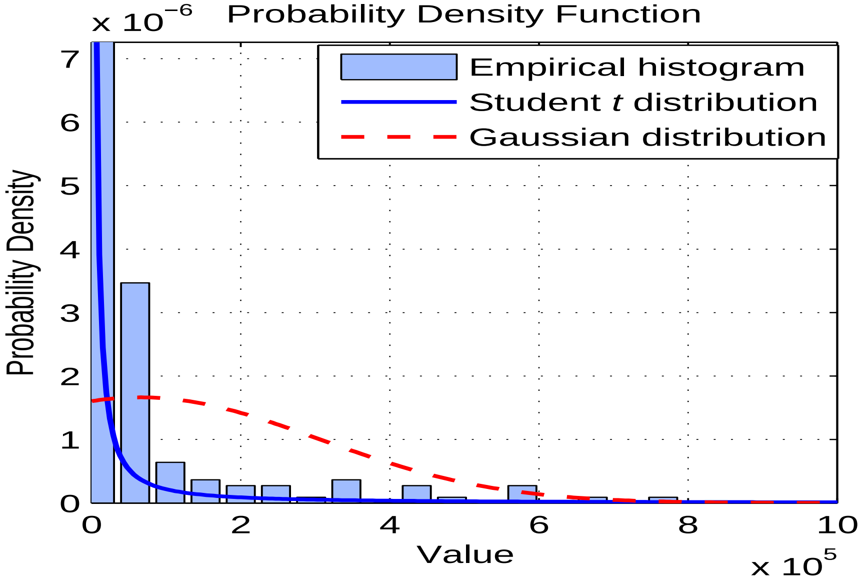


Fig. 7.6: Empirical histogram, Gaussian and Student t distribution of TE operated noise PWP coefficients at SNR of 15 dB

where

$$\chi_2 = \frac{\nu + 1}{2(\nu - 2)} \tag{7.5}$$

Following (7.1) to (7.4) in a similar way, student t pdf for $q_i(t_{k,m})$ can also be written as

$$q_i(t_{k,m}) = \frac{\chi_1}{\sigma_n}(1 - \chi_2 \frac{x^2}{\sigma_n^2}) \tag{7.6}$$

By substituting (7.4) and (7.6) in (3.6), we obtain

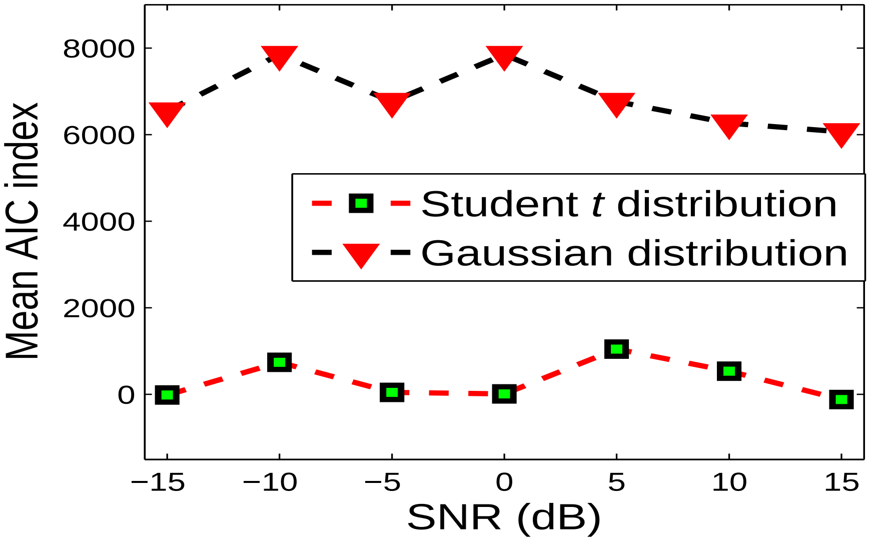


Fig. 7.7: Mean values of AIC index of TE operated PWP coeffiecients of noisy speech assuming Gaussian and Student t distributions

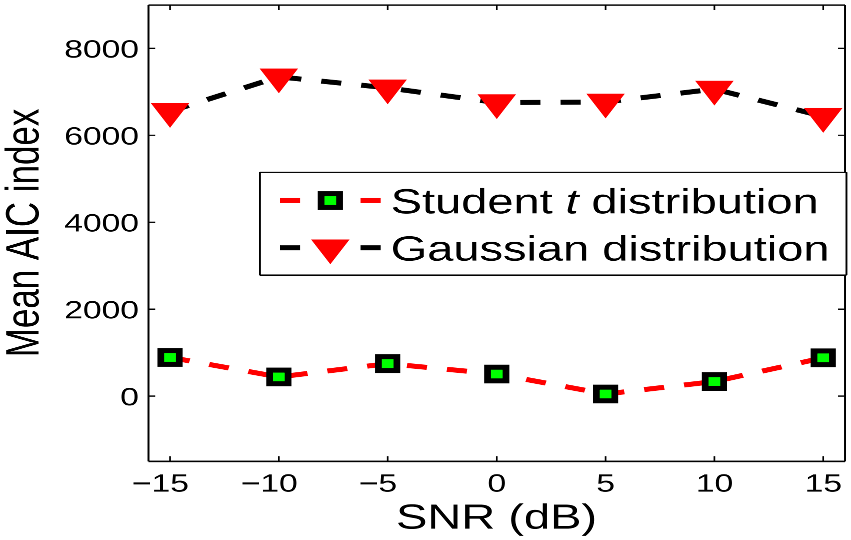


Fig. 7.8: Mean values of AIC index of TE operated PWP coeffiecients of noise assuming Gaussian and Student t distributions

$$\int_1^\lambda [\frac{\chi_1}{\sqrt{\sigma_r^2+\sigma_n^2}} - \frac{\chi_1\chi_2 x^2}{\sqrt{\sigma_r^2+\sigma_n^2}} - \frac{\chi_1}{\sigma_n} + \frac{\chi_1\chi_2 x^2}{\sigma_n^2})]I_1 dx = 0, \tag{7.7}$$

where $I_1 = ln(\frac{\sigma_n^2}{\sqrt{\sigma_r^2+\sigma_n^2}} \times \frac{1-\chi_1\frac{x^2}{\sigma_r^2+\sigma_n^2}}{1-\chi_1\frac{x^2}{\sigma_n^2}})$.
By solving (7.7), value of λ can be derived as

$$\lambda(k) = \sqrt{\frac{\sigma_n^2(k)(1+\gamma(k))}{(\sqrt{1+\gamma(k)}+2+\gamma(k))\sqrt{\chi_2}}}, \tag{7.8}$$

where $\gamma(k)$ is the segmental SNR of subband k defined as

$$\gamma(k) = \frac{\sigma_r^2(k)}{\sigma_n^2(k)}. \tag{7.9}$$

The proposed threshold $\lambda(k)$ in (7.8) derived assuming student t pdf is compared with that obtained assuming gaussian pdf given by

$$\lambda(k) = \frac{\sigma_n(k)}{\sqrt{\gamma_k}} \sqrt{2(\gamma_k + \gamma_k^2)} \times ln(\sqrt{1 + \frac{1}{\gamma_k}}) \tag{7.10}$$

in fig.7.9. This figure shows that the pattern of the threshold value is similar for both the pdfs at high as well as low SNRs. In terms of value, although student t pdf shows slightly lower values at high SNRs, but the threshold values are much lower than that of the gaussian pdf specially at low SNRs. Therefore, the threshold derived from the student t pdf offers less chance of removing speech coefficients while performing thresholding operation not only at high SNR but also at difficult low SNRs.

The proposed threshold as derived in (7.8) is high for higher noise power and low for lower noise power thus is adaptive to noise power of different subbands. In this method, voice activity detector is not needed as the threshold is automatically adapted to the silent and speech frames. At a silent frame, since noise power is significantly higher than the signal power, the proposed threshold results in a higher value as seen from (7.8). Such a value imposes more coefficients to be thresholded thus removing noise coefficients completely at subbands of a silent frame. Note that, in this paper, noise is estimated using Improved Minima Controlled Recursive Averaging (IMCRA) method [35].

7.0.3 Proposed Thresholding Function Considering Student *t* Statistical Model

We propose a Student t pdf dependent custom thresholding function derived from the modified hard and the semisoft thresholding functions [41]. Representing $\lambda(k)$ derived from (3.7) as $\lambda_1(k)$ and letting $\lambda_2(k) = 2\lambda_1(k)$, the proposed thresholding function is developed as in 3.9. In this thresholding function, shape parameters $\alpha(k, m)$ and $\beta(k, m)$ are determined assuming Student t distribution.

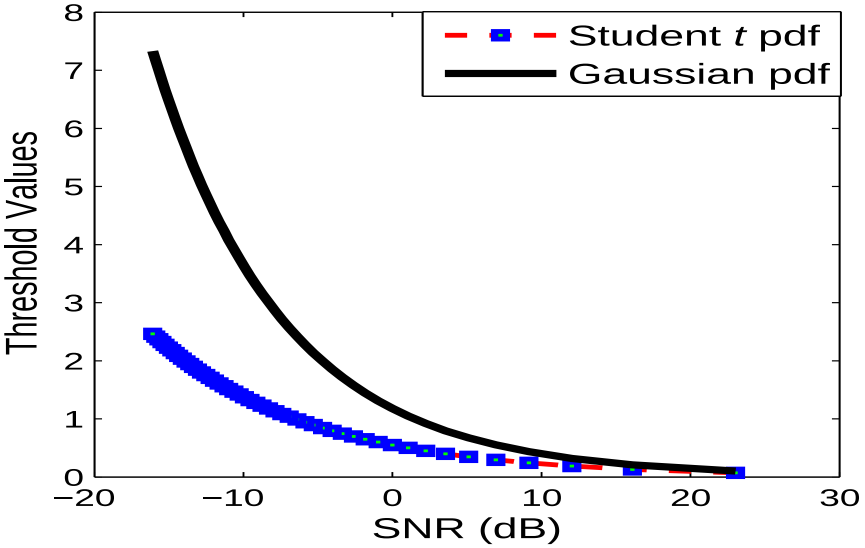


Fig. 7.9: Comparison of threshold values with respect to SNR for Student *t* and Gaussian pdfs

Determination of Shape Parameters assuming Student *t* distribution

The proposed thresholding function can be adapted to noise characteristics of the input noisy speech based on the shape parameters $\alpha(k,m)$ and $\beta(k,m)$ which are defined as

$$\alpha(k,m) = \frac{1+R(k,m)}{2(1+Q(k,m))}, \tag{7.11}$$

$$\beta(k,m) = \frac{2(1+Q(k,m))}{(1+R(k,m))}, \tag{7.12}$$

where $R(k,m)$ and $Q(k,m)$ are the speech presence and absence probabilities, respectively, of the m-th coefficient in the k-th subband and determined in the same method as used in [35].

Given two hypotheses,H_0 and H_1, which indicate respectively speech absence and presence in the m-th coefficient of the k-th subband, and assuming a student t distributions for both speech and noise PWP coefficients, the conditional pdfs of the speech and noise PWP coefficients are given by

$$f(Y(k,m)|H_0(k,m)) = \frac{\chi_1}{\sigma_n}(1+\frac{1}{\nu-2}\frac{Y(k,m)^2}{\sigma_n^2})^{-\frac{\nu+1}{2}} \tag{7.13}$$

$$f(Y(k,m)|H_1(k,m)) = \frac{\chi_1}{\sqrt{\sigma_r^2+\sigma_n^2}}(1+\frac{1}{\nu-2}\frac{Y(k,m)^2}{\sigma_r^2+\sigma_n^2})^{-\frac{\nu+1}{2}} \tag{7.14}$$

Using *aposteriori* and *apriori* SNRs defined by [6]

$$\Upsilon(k,m) = \frac{|Y(k,m)|^2}{\sigma_n^2(k,m)}, \tag{7.15}$$

$$\eta(k,m) = \frac{\sigma_r^2(k,m)}{\sigma_n^2(k,m)}, \tag{7.16}$$

and following (7.13) and (7.14), the conditional pdfs of the *aposteriori* SNR can be written as [35]

$$f(\Upsilon(k,m)|H_0(k,m)) = \frac{\chi_1}{\nu}(1 + \frac{\Upsilon(k,m)}{\nu - 2})^{-\frac{\nu+1}{2}} I_2 \tag{7.17}$$

$$f(\Upsilon(k,m)|H_1(k,m)) = \frac{\chi_1}{\nu\sqrt{(1+\eta(k,m))}}(1 + \frac{\Upsilon(k,m)}{(\nu-2)(1+\eta(k,m))})^{-\frac{\nu+1}{2}} I_2 \tag{7.18}$$

In (7.17) and (7.18), $I_2 = u(\Upsilon(k,m))$ is the unit step function. Noting that the conditional speech presence probability $R(k,m) = P(H_1(k,m)|\Upsilon(k,m))$, applying Bayes rule and using (7.18), an expression for $R(k,m)$ can be derived as

$$R(k,m) = [1 + \frac{Q(k,m)}{1 - Q(k,m)}(\sqrt{1 + \widehat{\eta}(k,m)})v(k,m)^{\frac{\nu+1}{2}}]^{-1}, \tag{7.19}$$

where $\widehat{\eta}(k,m)$ is the estimated *apriori* SNR obtained as in [35] and

$$v(k,m) = (1 + \frac{\Upsilon(k,m)}{\nu - 2})^{-1}(1 + \frac{\Upsilon(k,m)}{(\nu-2)(1+\widehat{\eta})}), \tag{7.20}$$

Speech absence probability $Q(k,m)$ in (7.19) can be determined as

$$Q(k,m) = 1 - R_{local}(k,m)R_{global}(k,m)R_{subband}(k,m), \tag{7.21}$$

In (7.21), $R_{local}(k,m)$ and $R_{global}(k,m)$ are the speech presence probabilities in local and global windows in the PWP domain. Letting τ for representing either "local" or "global" window, $R_\tau(k,m)$ can be given by

$$R_\tau(k,m) = \begin{cases} 0, & \text{if } \xi_\tau(k,m) \leq \xi_{min} \\ 1, & \xi_\tau(k,m) \geq \xi_{max}, \\ \frac{log(\xi_\tau(k,m)/\xi_{min})}{log(\xi_{max}/\xi_{min})}, & \text{otherwise} \end{cases} \tag{7.22}$$

where $\xi_\tau(k,m)$ representing either "local" or "global" average of the *apriori* SNR given by

$$\xi_\tau(k,m) = \sum_{i=-W_\tau}^{i=W_\tau} h_\tau(i)\xi(k-i,m) \tag{7.23}$$

In (7.23), h_τ is a normalized window of size $2w_\tau + 1$ and $\xi(k,m)$ represents a recursive average of the *apriori* SNR given by

$$\xi(k,m) = \kappa\xi(k,m-1) + (1-\kappa)\widehat{\eta}(k,m-1) \tag{7.24}$$

where κ denotes a smoothing constant. Note that in (7.22), ξ_{min} and ξ_{max} are the two empirical constants representing minimum and maximum values of $\xi(k,m)$ given in (7.24). $R_{subband}(k)$ in (7.21) can be computed as

$$R_{subband}(k) = \begin{cases} 0, & \text{if } \xi_{subband}(k) < \xi_{min} \\ 1, & \text{if } \xi_{subband}(k) > \xi_{subband}(k-1) and \xi_{subband}(k) > \xi_{min}, \\ \mu(k), & \text{otherwise} \end{cases} \tag{7.25}$$

where $\mu(k)$ is expressed as

$$\mu(k) = \begin{cases} 0, & \text{if } \xi_{subband}(k) \leq \xi_{peak}(k)\xi_{min} \\ 1, & \text{if } \xi_{subband}(k) \geq \xi_{peak}(k)\xi_{max}, \\ \frac{log(\xi_{subband}(k)/\xi_{peak}(k)/\xi_{min})}{log(\xi_{max}/\xi_{min})}, & \text{otherwise} \end{cases} \tag{7.26}$$

In (7.25) and (7.26), $\xi_{subband}(k)$ is determined as

$$\xi_{subband}(k) = \frac{1}{N_c} \sum_{1 \ll m \ll N_c} \xi(k,m) \tag{7.27}$$

and ξ_{peak} in (7.26) is a confined peak value of $\xi_{subband}(k)$. Thus computing $R(k,m)$ and $Q(k,m)$ following (7.19) and (7.21), the shape parameters $\alpha(k,m)$ and $\beta(k,m)$ can be determined using (7.11) and (7.12), respectively.

The enhanced speech frame is synthesized by performing the inverse PWP transformation PWP^{-1} on the resulting thresholded PWP coefficients and The final enhanced speech signal is reconstructed by using the standard overlap-and-add method.

7.1 Results Considering Student *t* Statistical Model

In this Section, a number of simulations is carried out with the same simulation conditions as described in chapter 3 to evaluate the performance of the proposed method considering Student t statistical model. Same comparison metrics are used to compare the proposed method with the previously mentioned comparison methods.

7.1.1 Simulation Conditions

Real speech sentences from the NOIZEUS database are employed for the experiments, where the speech data is sampled at 8 KHz [42]. To imitate a noisy environment, noise sequence is added to the clean speech samples at different SNR levels ranging from 15 dB to -15 dB. As in [43], two different types of noises, such as car and multi-talker babble are adopted from the NOIZEUS databases [42].

In order to obtain overlapping analysis frames, hamming windowing operation is performed, where the size of each of the frame is 64 ms (512 samples) with 50% overlap between successive frames. We get motivated to use 64 ms frame following the papers in [50] and [40]. A 6-level PWP decomposition tree with 10 db bases function is applied on the noisy speech frames resulting in subbands $k = 1, 2, ...24$ [38], [40]. The values of used constants to determine the shape parameters in the proposed thresholding function are given in Table 7.1.

We have tested our proposed method in a wide range of SNRs and reported the results in the SNR range of $15dB$ to $-15dB$, where a significant difference in performance is noticed for the proposed method relative to the other comparison methods. Our main focus was to show the capability of the proposed method at very low SNR levels, such as $-15dB$, where the other comparison methods produce less accurate results but the proposed method successfully enhances the speech with higher accuracy. On the other hand, in case of very high SNR, such as above $15dB$, although the proposed method consistently demonstrates better performance but the performance becomes competitive with respect to the other comparison methods. Therefore, the range of SNR used to present the comparative performance analysis is chosen from $15dB$ to $-15dB$. The parameters in Table I are selected empirically following [35].

Table 7.1: Constants used to determine the shape parameters

Constants	Value of constants
β	0.7
ξ_{min}	-10 dB
ξ_{max}	-5 dB
ξ_{peak}	10 dB
w_{local}	1
w_{global}	15

7.1.2 Comparison Metrics

Standard Objective metrics namely, Segmental SNR (SNRSeg) improvement in dB, Perceptual Evaluation of Speech Quality (PESQ) and Weighted Spectral Slope (WSS) are used for the evaluation of the proposed method [42]. The proposed method is subjectively evaluated in terms of the spectrogram representations of the clean, noisy and enhanced speech signals. Formal listening tests are also carried out in order to find the analogy between the objective metrics and the subjective sound quality. The performance of our method is compared with some of the state-of-the-art speech enhancement methods, such as Universal [10] and SMPO [43] in both objective and subjective senses. In SMPO method, speech is segmented into 20 ms frames and Han-windowed with 50% overlap. We have implemented the methods in [10] and [43] independently using the parameters specified therein. The implementation codes for [10] and [43] are obtained from very authentic publicly available sources. The Matlab code for [43] has been acquired from http://ecs.utdallas.edu/loizou/cimplants/ and the Matlab code for [10] developed by MATLAB Inc. has been used. The used built in function of MATLAB is given in http://www.mathworks.com/help/wavelet/ref/wden.html.

7.1.3 Objective Evaluation

Results for Speech signals with Car Noise

SNRSeg improvement, PESQ and WSS for speech signals corrupted with car noise for Universal, SMPO and proposed methods are shown in Fig. 7.10, Table 7.2 and Fig. 7.11.

In Fig. 7.10, the performance of the proposed method is compared with that of

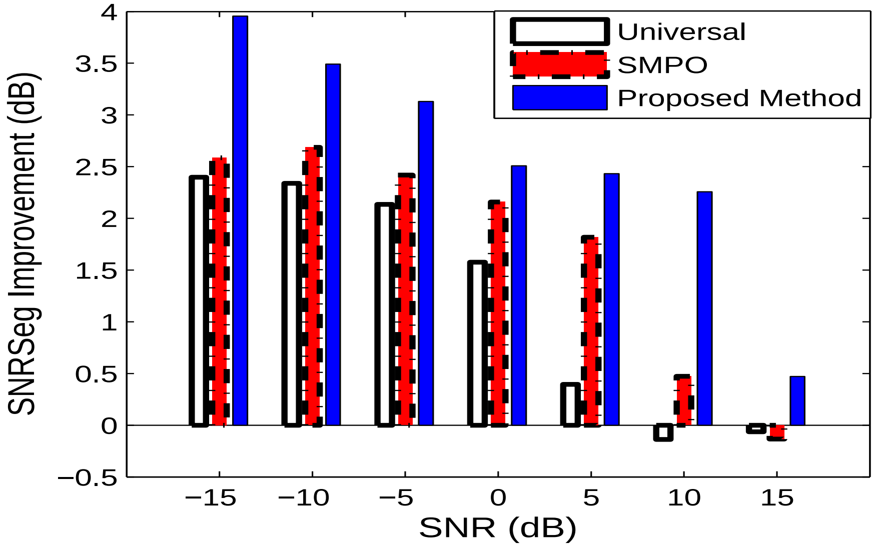


Fig. 7.10: SNRSeg Improvement for different methods in car noise

the other methods at different levels of SNR for car noise in terms of Segmental SNR inprovement. We see, the SNRSeg improvement increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields the highest SNRSeg improvement. Such larger values of SNRSeg improvement at a low level of SNR attest the capability of the proposed method in producing enhanced speech with better quality for speech severely corrupted by car noise.

In Table 7.2, it can be seen that at a low level of SNR, such as $-15dB$, all the methods show lower values of PESQ scores, whereas the PESQ score is much higher, as expected, for the proposed method. The proposed method also yields larger PESQ scores compared to that of the other methods at higher levels of SNR. Since, at a particular SNR, a higher PESQ score indicates a better speech quality, the proposed method is indeed better in performance in the presence of a car noise. For the same noisy conditions as in Table 7.2, we have also evaluated the PESQ results for the proposed and other two comparison methods using 20 ms frame size and hamming window. It is found that the proposed method is also better in performance while using 20 ms frame size and hamming window in the presence of a car noise.

Fig. 7.11 represents the WSS values as a function of SNR for the proposed method and that for the other methods. As shown in the figure, the WSS values resulting from all other methods are relatively larger for a wide range of SNR levels, whereas the proposed method is capable of producing enhanced speech with better quality as it gives lower values of WSS even at a low SNR of $-15dB$.

Table 7.2: PESQ for different methods in car noise

SNR(dB)	**Universal**	**SMPO**	**Proposed Method**
-15	1.16	1.15	1.40
-10	1.23	1.37	1.42
-5	1.32	1.51	1.80
0	1.43	1.69	1.97
5	1.65	2.07	2.28
10	1.93	2.38	2.71
15	2.14	2.60	2.96

In particular, for SMPO method, we have evaluated not only PESQ, but also other two objective parameters, SNRSeg improvement and WSS in the presence of car noise using both 20 ms and 64 ms frame size and hamming window. Comparing the PESQ results using 20 ms and 64 ms, it is found that PESQ results for SMPO is worse in the later case. But the increased size of frame improves the other two objective parameters, namely SNRSeg improvement and WSS for SMPO.

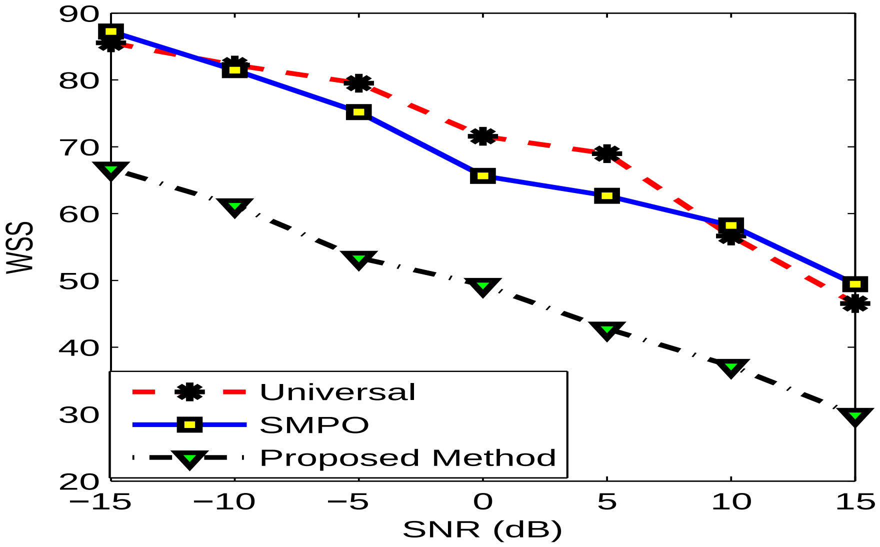


Fig. 7.11: WSS for different methods in car noise

Results for Speech signals with Multi-talker Babble Noise

SNRSeg improvement, PESQ and WSS for speech signals corrupted with babble noise for Universal, SMPO and proposed methods are shown in Fig. 7.12, 7.14 and 7.13, respectively.

In Fig. 7.12, it can be seen that at a low level of SNR of $-15dB$, the proposed method provides a SNRSeg improvement that is significantly higher than that of the methods of comparison. The proposed method still shows better performance in terms of SNRSeg improvement for higher SNRs also.

For speech corrupted with babble noise, in Fig. 7.13, the mean values of PESQ with standard deviation obtained using the proposed method is plotted and compared with that of the other methods. From this plot, it is seen that over the whole SNR range considered, the proposed method continue to provide higher PESQ with almost non-overlapping standard deviation in the presence of babble noise.

The performance of the proposed method is compared with that of the other methods in terms of WSS in Fig. 7.14 at different levels of SNRs in presence of babble noise. It is clearly seen from this figure that WSS increases as SNR decreases. At a low SNR of $-15dB$, the proposed method yields a WSS that is significantly lower than that of all other methods, which remains lower over the higher SNRs also.

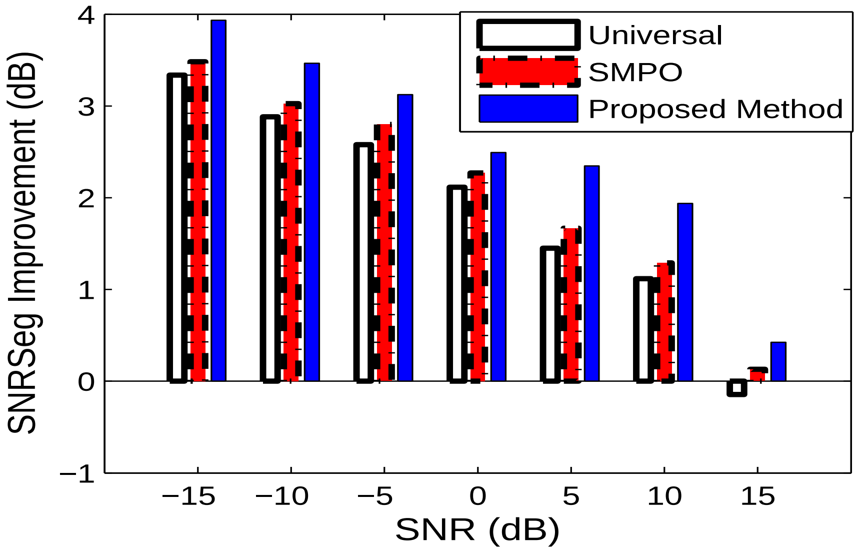


Fig. 7.12: SNRSeg Improvement for different methods in babble noise

7.1.4 Subjective Evaluation

In order to evaluate the subjective observation of the enhanced speech, spectrograms of the clean speech, the noisy speech, and the enhanced speech signals obtained by using the proposed method and all other methods are presented in Fig. 7.15 for car noise corrupted speech at an SNR of 10 dB. It is evident from this figure that the harmonics are well preserved and the amount of distortion is greatly reduced in the proposed method. Thus, the spectrogram observations with lower distortion also validate our claim of better speech quality as obtained in our objective evaluations

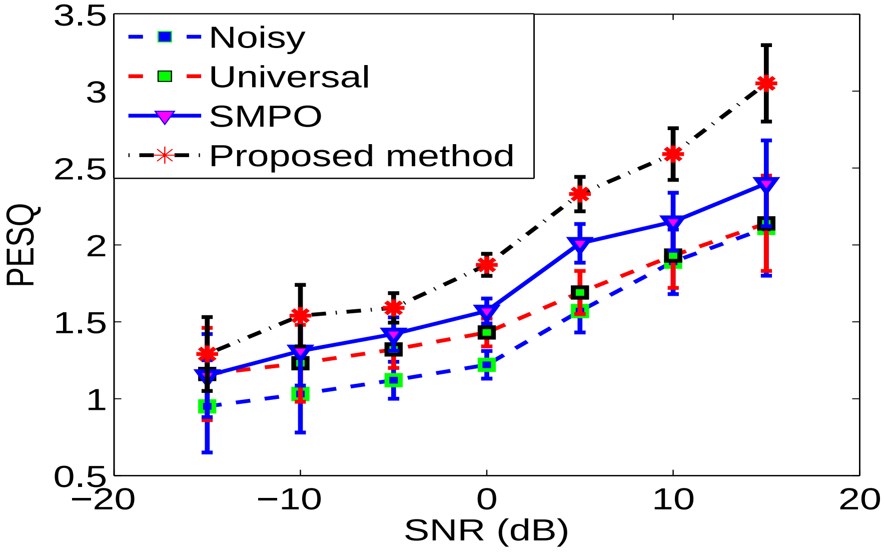


Fig. 7.13: PESQ for different methods in babble noise

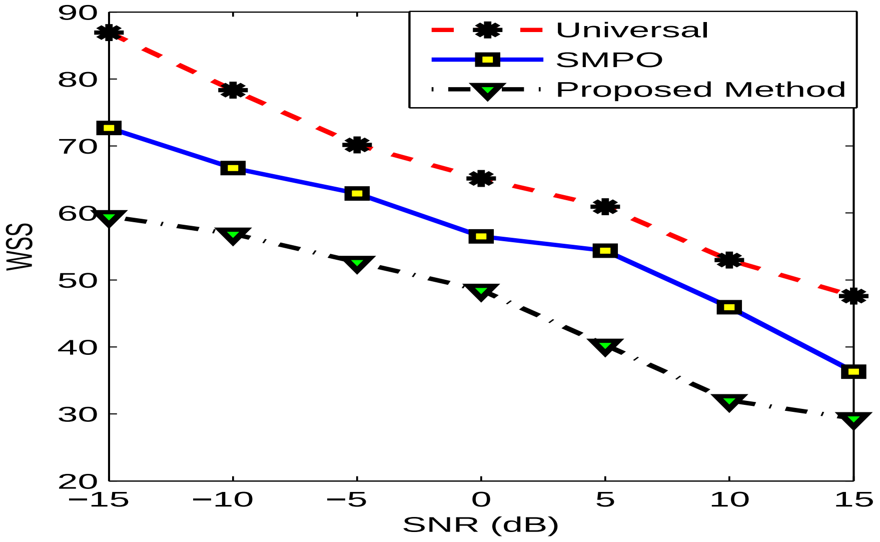


Fig. 7.14: WSS for different methods in babble noise

in terms of higher SNR improvement in dB, higher PESQ score and lower WSS in comparison to the other methods. Another set of spectrograms for babble noise corrupted speech at an SNR of 10 dB is also presented in Fig. 7.16. This figure attests that the proposed method has a better efficacy in preserving speech harmonics even in case of babble noise.

Extensive simulations have been carried out and it is seen that proposed method is capable of preserving the unvoiced or weak speech frames for most of the speech files of NOIZEUS database in the presence of noises at different SNR levels. The spectrograms for car and babble noises at $0dB$ and $-10dB$ are also analyzed and it is found that the proposed method outperforms the other methods both in removing noise and preserving the speech quality.

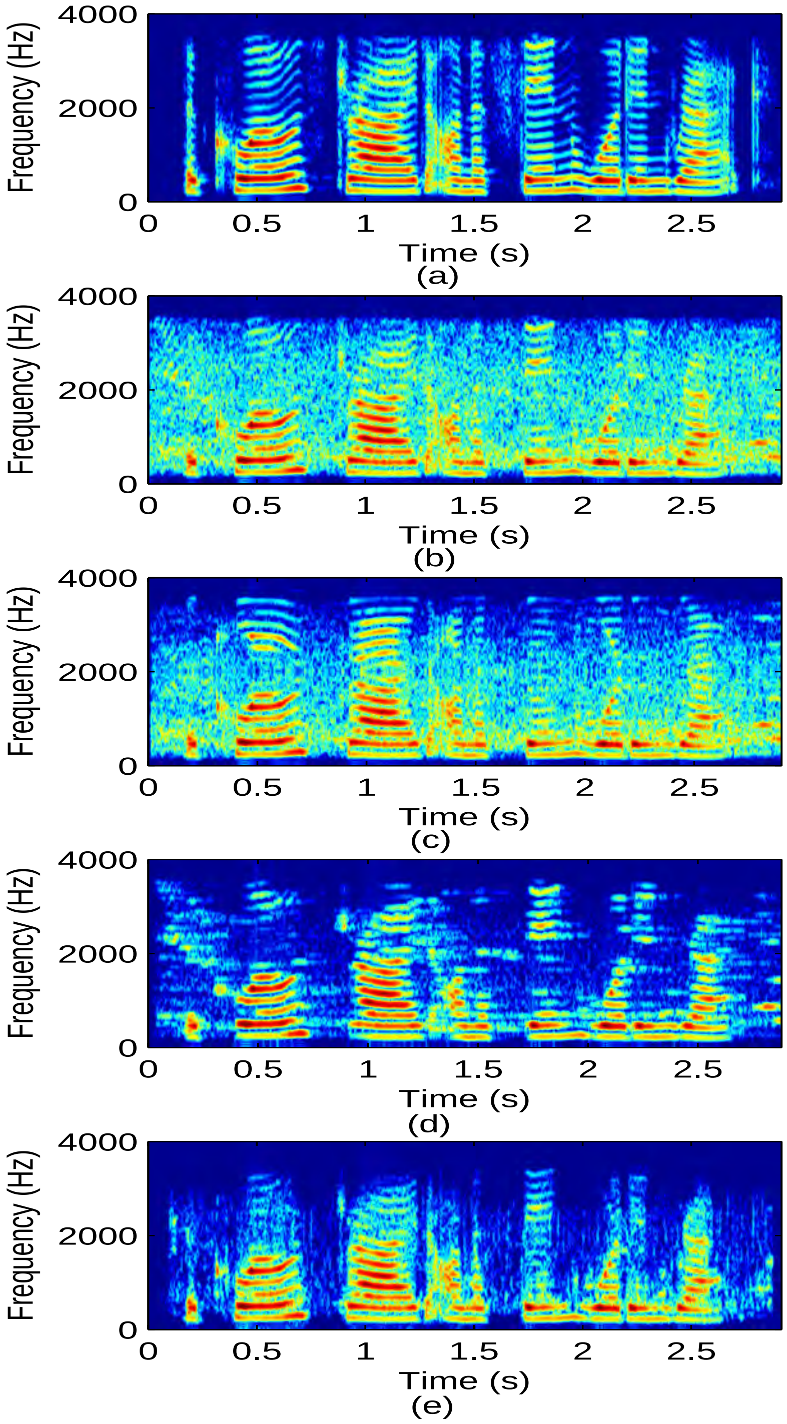


Fig. 7.15: Spectrograms of (a) Clean Signal (b) Noisy Signal with 10dB car noise; spectrograms of enhanced speech from (c) Universal method (d) SMPO method (e) Proposed method

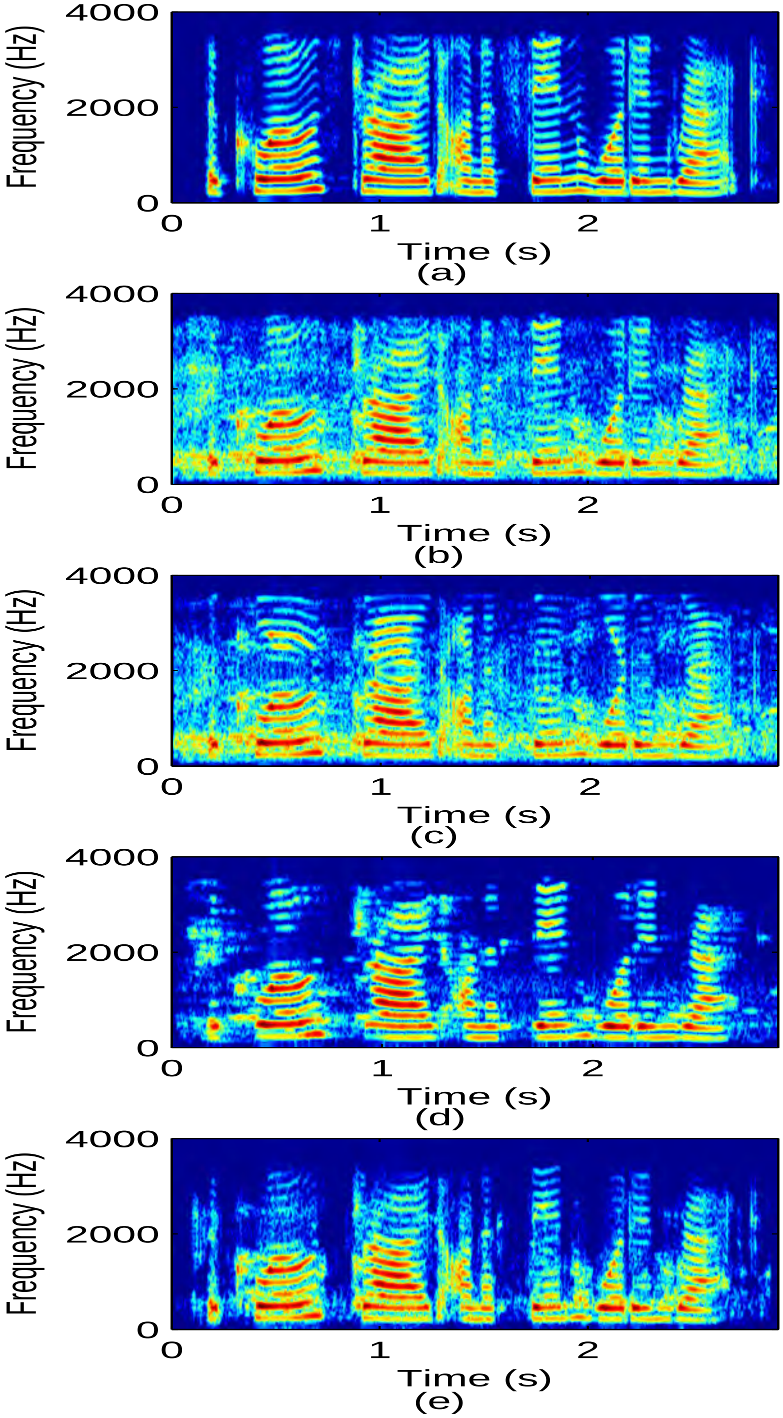


Fig. 7.16: Spectrograms of (a) Clean Signal (b) Noisy Signal with 10dB babble noise; spectrograms of enhanced speech from (c) Universal method (d) SMPO method (e) Proposed method

Formal listening tests are also conducted, where ten listeners are allowed and arranged to perceptually evaluate the enhanced speech signals. A full set (thirty sentences) of the NOIZEUS corpus was processed by Universal, SMPO and proposed method for subjective evaluation at different SNRs. Subjective tests were performed according to ITU-T recommendation P.835 [42]. In this tests, a listener is instructed to successively attend and rate the enhanced speech signal based on (a) the speech signal alone using a scale of SIG (1 = very unnatural, 5 = very natural), (b) the background noise alone using a scale of background conspicuous/ intrusiveness (BAK) (1 = very conspicuous, very intrusive; 5 = not noticeable), and (c) the overall effect using the scale of the mean opinion score (OVRL) (1 = bad, 5 = excellent). More details about the testing methodology can be found in [44]. The mean scores of SIG, BAK, and OVRL scales for the three speech enhancement methods evaluated in the presence of car noise at an SNR of 5 dB are shown in Table 7.3. For the three methods evaluated using babble noise-corrupted speech at an SNR of 5 dB, the mean scores of SIG, BAK, and OVRL scales are also summarized in Table 7.4. The mean scores in the presence of both car and babble noises demonstrate that the lower signal distortion (i.e., higher SIG scores) and the lower noise distortion (i.e., higher BAK scores) are obtained with the proposed method relative to that obtained by Universal and SMPO methods in most of the conditions. It is also shown that a consistently better performance in OVRL scale is offered by the proposed method not only in car but also in babble noisy conditions in comparison to that provided by all the methods mentioned above. Overall, it is found that the proposed method possesses the highest subjective sound quality in comparison to that of the other methods in case of different noises.

7.2 Conclusions

In this paper, we developed a Student t statistical model-based technique for the TE operated PWP coefficients of the noisy speech in order to obtain a suitable threshold value. Unlike the unique threshold based method, the threshold value thus obtained is adaptive in nature based on the speech and silence subbands. By employing the proposed Student t pdf dependent custom thresholding function, the PWP coefficients of the noisy speech are thresholded in order to obtain an enhanced

Table 7.3: Mean Scores of SIG, BAK and OVL scales for different methods in presence of car noise at 5 db

Listener	Universal	SMPO	Proposed Method
1	2.6	4.0	4.1
2	3.3	3.8	3.7
3	3.9	4.1	4.3
4	3.6	4.2	4.2
5	3.3	3.9	4.1
6	3.9	4.6	4.9
7	3.8	3.8	4.3
8	3.6	4.1	4.2
9	3.5	4.5	4.7
10	3.9	4.8	4.9

Listener	Universal	SMPO	Proposed Method
1	4.0	4.5	5.0
2	4.3	4.9	4.7
3	4.2	4.4	4.9
4	4.4	4.7	4.8
5	4.2	4.8	4.7
6	3.9	4.6	4.9
7	3.8	3.9	4.4
8	4.4	4.6	4.6
9	3.5	3.8	4.5
10	4.2	4.5	4.8

Listener	Universal	SMPO	Proposed Method
1	3.6	4.0	4.0
2	3.3	3.9	3.7
3	3.9	4.0	4.2
4	3.4	4.2	4.5
5	3.2	3.8	4.0
6	2.9	3.6	3.9
7	3.8	3.8	4.2
8	3.5	3.7	4.2
9	3.5	3.9	3.8
10	3.7	3.9	4.0

Table 7.4: Mean Scores of SIG, BAK and OVL scales for different methods in presence of babble noise at 5 db

Listener	Universal	SMPO	Proposed Method
1	2.9	4.1	4.3
2	3.6	3.9	3.9
3	3.8	4.3	4.2
4	3.7	4.0	4.1
5	3.6	3.8	4.4
6	3.8	4.4	4.8
7	3.9	3.9	4.4
8	3.6	4.1	4.3
9	3.7	4.4	4.8
10	3.8	4.6	4.8

Listener	Universal	SMPO	Proposed Method
1	3.1	4.0	4.2
2	3.5	3.7	3.8
3	3.8	4.2	4.4
4	3.4	4.1	4.3
5	3.5	3.8	4.3
6	3.7	4.5	4.8
7	3.9	3.9	4.6
8	3.8	4.3	4.4
9	3.7	4.4	4.8
10	3.9	4.7	4.8

Listener	Universal	SMPO	Proposed Method
1	3.6	4.2	4.3
2	3.5	3.9	3.9
3	3.5	4.3	4.4
4	3.6	4.3	4.4
5	3.5	3.8	4.2
6	3.8	4.5	4.8
7	3.7	3.9	4.1
8	3.7	4.2	4.4
9	3.6	4.4	4.6
10	3.8	4.7	4.8

speech. Simulation results show that the proposed method yields consistently better results in the sense of higher Segmental SNR Improvement in dB, higher output PESQ, and lower WSS values than those of the existing methods. The improved performance of the proposed method is also attested by the much better spectrogram outputs and in terms of the higher scores in the formal subjective listening tests.

Chapter 8

Conclusion

8.1 Concluding Remarks

An improved perceptual wavelet packet transform based approach to solve the problems of speech enhancement using the Probability distributions of Teager Energy Operated perceptual wavelet Packet coefficients has been presented in this paper. We incorporated a statistical model-based techniques with teager energy operation on the of the perceptual wavelet packet coefficients to obtain a suitable adaptive threshold using symmetric K-L divergence. We also design custom thresholding functions to provide better speech enhancement.

8.2 Contribution of the Thesis

The major contributions of this thesis are:

1. Statistical models for determining an adaptive threshold is proposed using Gaussian, Laplace, Rayleigh, Poisson and Student t distribution functions of the TE operated perceptual wavelet packet coefficients.

2. Custom thresholding functions are proposed that combine different thresholding techniques and able to provide better thresholding than the thresholding functions described as in the literature.

3. Detail simulations have been carried out in order to investigate the performance of the proposed methods in terms of objective and subjective senses.

4. The performance of our proposed methods is compared with state-of-the-art methods, namely Universal and SMPO.

5. Simulation results show that the proposed methods yield consistently better results in the sense of higher output segmental SNR improvement, in dB, higher output PESQ, and lower WSS values than those of the existing methods. The proposed methods are also found consistently better in spectrogram observations and formal listening tests.

8.3 Scopes for Future Work

However, there are still some scopes for future research, as mentioned below:

1. Available databases other than NOIZEUS may be utilized for testing the efficacy of our proposed methods.

2. The IMCRA method of noise estimation is used in all our methods. A better noise estimation can be exploited to obtain more effective performances.

Bibliography

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice.* Boca Raton: CRC Press, 2007.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, 1979.

[3] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. IV–4164, 13–17.

[4] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 251–266, 1995.

[5] B. Chen and P. C. Loizou, "A laplacian-based (mmse) estimator for speech enhancement," *Speech Communication*, vol. 49, pp. 134–143, 2007.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 443–445, 1985.

[7] I. Almajai and B. Milner, "Visually derived wiener filters for speech enhancement," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 1642–1651, 2011.

[8] S. Tabibian, A. Akbari, and B. Nasersharif, "A new wavelet thresholding method for speech enhancement based on symmetric kullback-leibler diver-

gence," in *14th International CSI Computer Conference (CSICC)*, 2009, pp. 495–500.

[9] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the teager energy operator," *IEEE Signal Processing Letters*, vol. 8, pp. 10–12, 2001.

[10] D. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, pp. 613–627, 1995.

[11] H. Sheikhzadeh and H. R. Abutalebi, "An improved wavelet-based speech enhancement system," *EUROSPEECH*, pp. 1855–1858, 2001.

[12] Y. Shao and C. H. Chang, "A generalized timefrequency subtraction method for robust speech enhancement based on wavelet filter banks modeling of human auditory system," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 37, pp. 877–889, 2007.

[13] D. A. Reynolds and R. C. Rose, "Robust text-dependent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, 1995.

[14] Y. X. Johnson, M. T. and Y. Ren, "Speech signal enhancement through adaptive wavelet thresholding," *Speech Communication*, 2007.

[15] Y. Ghanbari and M. Mollaei, "A new approach for speech enhancement based on the adaptive thresholding of the wavelet packets," *Speech Communication*, vol. 48, pp. 927–940, 2006.

[16] Y. B. Chang, S. G. and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Transaction on Image Processing*, vol. 9, pp. 1532–1546, 2000.

[17] B. T. Luisier, F. and M. Unser, "A new sure approach to image denoising: Interscale orthonormal wavelet thresholding," *IEEE Transaction on Image Processing*, vol. 16, pp. 593–606, 2007.

[18] D. OShaughnessy, *Speech Communications: Human and Machine.* Wiley-IEEE Press, 1999.

[19] J. H. J. Jr. Deller and J. Proakis, *Discrete-Time Processing of Speech Signals.* NY: IEEE Press, 2000.

[20] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," in *Proc. IEEE*, Dec. 1979, pp. 221–239.

[21] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 126–137, 1999.

[22] U. Mittal and N. Phamdo, "Signal/noise klt based approach for enhancing speech degraded by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 159–167, 2000.

[23] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 334–341, 2003.

[24] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 700–708, 2003.

[25] K. S. N. You, C. H. and S. Rahardja, "An invertible frequency eigen domain transformation for masking-based subspace speech enhancement," *IEEE Signal Processing Letters*, vol. 12, pp. 461–464, 2005.

[26] N. S. Gustafsson, H. and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 799–807, 2001.

[27] K. Yamashita and T. Shimamura, "Nonstationary noise estimation using low-frequency regions for spectral subtraction," *Signal Processing Letters*, vol. 12, pp. 465–468, 2005.

[28] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 857–869, 2005.

[29] R. V. Hansen, J. H. L. and K. Arehart, “Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 2049–2063, 2006.

[30] R. Martin, “Speech enhancement based on minimum mean-square error estimation and super gaussian priors,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 845– 856, 2005.

[31] S. Ben Jebara, “A perceptual approach to reduce musical noise phenomenon with wiener denoising technique,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2006, pp. 14–19.

[32] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, 2002.

[33] K. Y. Y. S. I. Chang, S. and I. J. Kim, “Speech enhancement for non-stationary noise environment by adaptive wavelet packet,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. I–561 – I–564.

[34] H. Yi and P. Loizou, “Speech enhancement based on wavelet thresholding the multitaper spectrum,” *IEEE Signal Processing Letters*, vol. 12, pp. 59–67, 2004.

[35] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Transactions on Speech and Audio Processing*, vol. 11, 2003.

[36] Q. Fu and E. A. Wan, “A novel speech enhancement system based on wavelet denoising,” *Center of Spoken Language Under standing, OGI School of Science and Engineering at OHSU*, 2003.

[37] M. Islam and C. Shahnaz, “Enhancement of noisy speech based on a threshold determined through Gaussian modeling and a PDF dependent thresholding function,” *accepted with revision in EURASIP Journal on Audio, Speech, and Music Processing, ID: MS:2128583861707195*, 2015.

[38] P. B. H. J. Sarikaya, R., “Wavelet packet transform features with application to speaker identification,” 1998.

[39] J. Kaiser, “Some useful properties of teagers energy operators,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 149–152.

[40] T. Sanam and C. Shahnaz, “Noisy speech enhancement based on an adaptive threshold and a modified thresholding function in wavelet packet domain,” *Digital Signal Processing*, vol. 23, pp. 941–951, 2013.

[41] ——, “Enhancement of noisy speech based on a custom thresholding function with a statistically determined threshold,” *International Journal of Speech Technology*, vol. 15, pp. 463–475, 2012.

[42] Y. Hu and P. Loizou, “Subjective comparison and evaluation of speech enhancement algorithms,” *Speech Communication*, vol. 49, pp. 588–601, 2007.

[43] Y. Lu and P. Loizou, “Estimators of the magnitude-squared spectrum and methods for incorporating snr uncertainty,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1123–1137, 2011.

[44] ITU, “P835 it: subjective test methodology for evaluating speech communication systems that include noise suppression algorithms.” *ITU-T Recommendation (ITU, Geneva)*, p. 835, 2003.

[45] M. Islam and C. Shahnaz, “Speech enhancement in the perceptual wavelet packet domain based on a threshold derived from Laplace modeling,” *Under review in Journal of the Acoustical Society of America, ID: MS 15-15379*, 2014.

[46] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol. 19, pp. 716–723, 1974.

[47] M. Islam and C. Shahnaz, “Rayleigh modeling of Teager Energy operated perceptual wavelet packet coefficients for enhancing noisy speech,” *accepted with revision in Speech Communication, Elsevier, ID: Ms. No. SPECOM-D-15-00031*, 2015.

[48] ——, "Modeling of Teager Energy operated perceptual wavelet packet coefficients based on Poisson distribution for noisy speech enhancement," *To be submitted to IET Signal Processing*, 2015.

[49] ——, "Speech enhancement based on student t modeling of Teager Energy operated perceptual wavelet packet coefficients and a PDF dependent thresholding function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 1800 – 1811, 2015.

[50] S. Ganapathy and J. Pelecanos, "Enhancing frequency shifted speech signals in single side-band communication," *IEEE Signal Process. Lett.*, vol. 20, pp. 1231–1234, 2013.